# Dativalternering i norsk

## *En probabilistisk tilnærming*

Kjell Gunnar Styve


Masteroppgave ved ILN/HF

UNIVERSITETET I OSLO

26.04.2012

# Dativalternering i norsk

## *En probabilistisk tilnærming*

Kjell Gunnar Styve

Masteroppgave ved ILN/HF

UNIVERSITETET I OSLO

26.04.2012

2012

Dativalternering i norsk: en probabilistisk tilnærming

Kjell Gunnar Styve

http://www.duo.uio.no/

Trykk: Reprosentralen, Universitetet i Oslo

# Sammendrag

Hva er det som avgjør om man sier ’Hun ga mannen en penn’ eller ’Hun ga en penn til mannen’? Fenomenet kalles dativalternering, og tilsynelatende er det fri veksling mellom disse to konstruksjonstypene.

I løpet av de siste ti årene har Joan Bresnan og ulike medforfattere publisert et antall artikler om dativalternering i engelsk der hun legger fram konkrete probabilistiske modeller, bygget på et elektronisk korpus, for valg av konstruksjonstype. Disse modellene viser at dette valget i engelsk i stor grad avhenger av bestemte egenskaper ved recipient og theme, som for eksempel hvorvidt de er kontekstuelt gitt, om de er definitte, om de er pronominale, om de er animate, og deres relative kompleksitet målt i fraselengde. Denne masteroppgaven søker å replikere dette arbeidet med norske data hentet fra ulike korpora tilgjengelig fra tekstlaboratoriet ved ILN/UiO.

Oppgaven legger hovedvekten på verbet *gi*, men legger også fram modeller bygget på et datasett med andre ditransitive verb. Modellene, såkalte logistiske regresjonsmodeller, er av en type som er mye brukt i andre vitenskapsgrener som undersøker multivariable sammenhenger. Oppgaven ville ikke vært mulig å gjennomføre uten tilgang til det statistiske programmet R, som egentlig er et komplett programmeringsspråk.

Modellene viser at talespråk oppfører seg noe ulikt skriftspråk, for så vidt som skriftlige ytringer er noe vanskeligere å predikere hva valg av konstruksjonstype angår. Modellene viser også at det generelt er lettere å predikere en dobbelt-objekt-konstruksjon V-NP-NP enn en preposisjonsfrase-konstruksjon V-NP-PP. Til sist viser modellene at verbet *gi* synes å være mindre predikabelt enn andre ditransitive verb hva valg av konstruksjonstype angår.

Teorien bak disse modellene bygger på en antakelse om at syntaksen gjenspeiler prominens på noen universelle prominenshierarkier. Forventningen er at konstituenter (recipient eller theme) med høy prominens på disse skalaene, blir realisert i en prominent posisjon syntaktisk, nemlig umiddelbart etter verbet. Resultatet av modelleringene synes å bekrefte denne forventningen, men viser også at det gjenstår en rest av tilsynelatende fri variabilitet i realiseringen. Det har ikke vært mulig innenfor rammen av denne oppgaven å undersøke om dette kan skyldes at andre faktorer også bidrar til valg av konstruksjonstype, eller om disse konstruksjonene i siste instans representerer ekte (frie) syntaktiske alternativer.

# Forord

Denne masteroppgaven er inspirert av Joan Bresnans mange arbeider inne rammen av probabilistisk syntaks gjennom de siste ti årene. Jeg søker å replikere hennes metoder på norske data, men lykkes dessverre ikke i samme grad. Om dette skyldes språklige forskjeller eller min anvendelse av metoden, har jeg ikke vært i stand til å avgjøre. Det gjenstår altså mye å gjøre på dette feltet i norsk, men jeg håper at oppgaven i det minste kan vise vei.

I stedet for å takke alle som har vært velvillige tilskuere eller hjelpere i dette arbeidet, vil jeg nøye meg med å konstatere som man gjør på mine hjemtrakter:

Ka ein enn gjere, so gjere ein da aot seg sjølvù.

# Innholdsfortegnelse

# 1 Innledning

Hva avgjør om man sier (1.1) eller (1.2)?

(1.1) Hun ga mannen en penn.

(1.2) Hun ga en penn til mannen.

Fenomenet kalles dativalternering og finnes blant annet i norsk og engelsk.

Haspelmath (2005, s. 426-427) oppgir i en typologisk oversikt over ditransitive konstruksjoner, begrenset til verbet 'gi', at disse kan realiseres på en av følgende måter:

- Indirekte-objekt-konstruksjoner, der theme har samme markering som patiens har i vanlige monotransitive konstruksjoner, mens recipienten avviker gjennom kasus eller adposisjon.
- Dobbelt-objekt-konstruksjoner, der theme og recipient har samme markering som patiens har i en vanlig monotransitiv konstruksjon.
- Sekundært-objekt-konstruksjoner, der recipienten har samme markering som patiens har i vanlige monotransitive konstruksjoner, mens theme avviker.
- Blandede konstruksjoner, det vil si at språket tillater minst to av de foregående konstruksjonene som alternative valg.

Dativalternering er et eksempel på blandede konstruksjoner, i og med at man da kan veksle mellom en indirekte-objekt-konstruksjon og en dobbelt-objekt-konstruksjon. I Haspelmaths utvalg bruker om lag en tiendedel av de totalt knapt 400 undersøkte språkene blandede konstruksjoner. I Europa bruker de aller fleste språkene bare dobbelt-objekt-konstruksjoner, mens dativalternering her i hovedsak er begrenset til germanske språk uten kasusmarkering (Primus, 1998, s. 440), det vil si engelsk, frisisk, nederlandsk, norsk, svensk og dansk. I tysk realiseres recipienten for eksempel vanligvis som et tradisjonelt indirekte objekt i dativ, mens theme realiseres som direkte objekt i akkusativ.

I norsk er spørsmålet altså om et ditransitivt verb gir opphav til en dobbeltobjekt-konstruksjon eller om recipienten realiseres i en preposisjonsfrase (med *til* eller *for*). I generativ lingvistikk har man vært mest opptatt av hvordan alternativene skal representeres, og hvordan ett av alternativene kan avledes (deriveres syntaktisk) fra det andre. I funksjonalistisk orientert lingvistikk har man også vært interessert i hvilke faktorer som påvirker valget av alternativ. I de siste årene har probabilistisk orienterte lingvister også søkt å kvantifisere de ulike faktorenes relative bidrag til valg av alternativ, blant annet ved å konstruere eksplisitte matematiske modeller for hvordan valg av alternativ realiseres i ulike varianter av engelsk, slik disse framtrer i relevante korpus. Denne oppgaven vil søke å gjøre det samme for norsk skrift- og talespråk.

I Norsk Referansegrammatikk (Faarlund et al, 2006) beskrives dativalterneringen i kapitlet ”Indirekte objekt”. Der påpekes det at ikke alle ditransitive verb kan ha denne alterneringen: noen verb kan bare ta vanlig indirekte objekt, ikke preposisjonsfrase med *til*. Verb som kan alternere er gjerne overføringsverb (*gi*, *sende*, *fortelle*, etc), som kan inndeles videre etter arten av det objektet som overføres: gjenstander, pengesummer, meningsinnhold, sanseinntrykk, etc. Alterneringen oppgis å være styrt av pragmatiske og referensielle forhold (som hvorvidt et ledd har unik referanse eller representerer kjent informasjon), lengden eller kompleksiteten til de to objektene, samt abstraksjonsgraden til det som overføres. Det går ikke fram hvordan disse ulike prinsippene for dativalternering antas å samvirke hvis de hver for seg trekker i ulike retninger. Denne oppgaven vil søke å belyse og i noen grad kvantifisere dette.

Det er neppe mulig å få full oversikt over alle faktorer som påvirker valg av alternativ i et konkret tilfelle. I denne oppgaven vil jeg se på et antall faktorer som har vært nevnt i litteraturen og som det er praktisk mulig å bestemme ut fra korpustreff. Dette dreier seg om semantiske forhold ved verbene som undersøkes, formelle trekk ved de to mulige objektene, samt deres relative fraselengde. Dette er forklart i detalj i senere kapitler. Det er selvfølgelig mulig at andre faktorer, som for eksempel stilistiske krav i skriftlig norsk, kan være like viktige som de som det her er tatt hensyn til. Og i siste instans gjenstår selvfølgelig den mulighet at valget av alternativ i stor grad er vilkårlig eller avhenger av individuelle preferanser, slik at et korpus som samler ytringer fra mange ulike kilder ikke er et relevant datagrunnlag. Jeg håper å kunne vise at dette ikke er tilfelle.

Oppgaven avhenger i vesentlig grad av matematisk modellering og etterfølgende analyse. Jeg vil søke å forklare dette så grundig som nødvendig for å kunne følge analysen, men heller ikke mer. Oppgaven er ikke ment som en lærebok i bruk av probabilistiske metoder i syntaks. Den er heller ikke ment som en lærebok i bruk av statistikkprogrammet R som jeg har brukt i modelleringen, men jeg vil likevel ta med nok detaljer om hvordan arbeidet rent praktisk er gjort til at det skal kunne replikeres eller tilpasses til andre typer syntaktiske alterneringer. Jeg håper å kunne vise at slike metoder og verktøy kan være nyttige ikke bare i anvendt lingvistikk.

En probabilistisk modell for valg av syntaktiske alternativer har åpenbare anvendelser innen datalingvistikk og NLP (natural language processing), men kan også bidra til at man får bedre oversikt over konkret hvordan ulike faktorer kan tenkes å samvirke innen teoretiske rammeverk som for eksempel stokastisk optimalitetsteori. Oppgaven vil derfor i noen grad også ta opp forholdet mellom denne analysen og slike beslekta teorier. Jeg håper å kunne vise at en probabilistisk modell av denne typen er minst like interessant som disse.

# 2 Litteraturgjennomgang

Fenomenet dativalternering har blitt forklart på ulike måter, som kan samles i tre hovedgrupper av forklaringer (Krifka 2003, Levin og Rappaport Hovav 2005): forklaringer som hevder at begge alternativene har samme betydning (”The Monosemy view” hos Krifka, ”Structure-driven analyses” hos Levin og Rappaport Hovav), forklaringer som hevder at alternativene har ulik betydning (”The Polysemy view” respektive ”Meaning-driven analyses”), og forklaringer som hevder at hensyn til informasjonsstruktur har avgjørende betydning (”The Information Structure view” respektive ”Information-packaging considerations”).

De tre forskjellige synene på dativalternering: monosemi, polysemi og informasjonsstruktur, gjennomgås kort i Krifka (2003, s. 1-3). I følge det monosemi-baserte synet har altså begge alternativene samme grunnleggende betydning. De er relatert til hverandre ved at ett av alternativene avledes syntaktisk av det andre, eller ved at samme argumentstruktur kan realiseres ved to forskjellige syntaktiske strukturer. Krifka påpeker at dette synet ser bort fra at dativalterering ikke gjelder alle ditransitive verb, noe som jo tyder på underliggende semantiske begrensninger.

Det polysemi-baserte synet skiller mellom to ulike betydninger: overføring av eierskap til theme eller skifte av lokasjon for theme. I følge dette synet er den grunnleggende betydningen av dobbeltobjekt-konstruksjonen

(2.1) Jenta ga gutten en bok

at eierskapet eller disposisjonsretten til den omtalte boka ble overført fra jenta til gutten. Den grunnleggende betydningen av preposisjonsfrase-konstruksjonen

(2.2) Jenta ga en bok til gutten

er derimot at boka ble forflyttet fra jenta til gutten. Krifka påpeker at betydningsforskjellen for noen verb kan være liten, noe som synes å være tilfelle i dette eksempelet. Han hevder videre at sannhetsbetingelsene ofte kan være sammenfallende, men at noen verb kan være kompatible med bare en av betydningene, slik at dette kan forklare hvorfor ikke alle ditransitive verb tar del i dativalternering

Det tredje synet, basert på informasjonsstruktur, mener Krifka kan være kompatibelt med både et monosemi- og polysemi-basert syn. Hvis det ikke er betydningsforskjell mellom alternativene, kan informasjonsstruktur meget vel være den avgjørende faktoren for valg av alternativ. På samme måte kan informasjonsstruktur i en gitt kontekst overstyre eventuelle minimale betydningsforskjeller mellom alternativene.

Det bredere problemet med det som kalles "multiple argument realization", inklusive dativalternering, gjennomgås i Levin og Rappaport Hovav (2005, kapittel 7). De konkluderer (s. 219), som Krifka, med at når betydningsforskjellen mellom alternativene er liten, slik de mener tilfellet er med dativalterneringen generelt, kan valg av alternativ brukes for informasjonspakkingsformål og for å ta hensyn til relativ vekt mellom argumentene. I denne boka refererer de struktur-drevne analyser, betydnings-drevne analyser og analyser basert på informasjonspakkingshensyn. Disse faller dermed sammen med Krifkas monosemi-, polysemi-, og informasjonsstruktur-baserte syn. I boka tar de ikke standpunkt til hvilke analyser som bør foretrekkes, men i Levin og Rappaport Hovav (2002), som bare behandler dativalternering, tar de eksplisitt standpunkt til fordel for et monosemi-basert syn der informasjonsstruktur får avgjørende betydning for valg av alternativ.

Forklaringer som hevder at begge alternativene har samme betydning kan være generative eller ikke-generative. Generative forklaringer regner ofte PP-varianten (*til*-varianten) som grunnleggende (Levin og Rappaport Hovav, 2005, s. 196), mens NP-varianten (dobbeltobjekt-konstruksjonen) er avledet av denne, eller følger av leksikalske regler. Ikke-generative forklaringer regner også ofte PP-varianten som grunnleggende (Levin og Rappaport Hovav, 2005, s. 202). Krifka (2003, s. 2) refererer flere ulike eksempler på slike analyser. I følge Larson (1988) er bruk av preposisjonsfrase den grunnleggende strukturen, og bruk av dobbelt objekt avledes av denne (som gjengitt i Krifka):

$$[_{v'}\ \text{give}_i[_{vp}\ \text{the car}[_{v'}\ t_i[_{pp}\ \text{to Beth}]]]] \Rightarrow [_{v}\text{give}_i[_{vp}\text{Beth}_j[_{v'}[_{v'}\ t_i\ t_j]\ \text{the car}]]]$$

Butt et al (1997), som arbeider innenfor LFG, opererer med følgende strukturer (tilpasset etter Krifka):

- Argumentstruktur: give(AGENT, RECIPIENT, THEME)
- Syntaktiske realiseringer:

1. give [Beth]OBJ [the car]$OBJ_{THEME}$
2. give [the car]OBJ [to Beth]$OBL_{RECIPIENT}$

(En slik analyse brukes ikke bare i LFG. Tallerman (2005, s. 180) nevner at det i engelsk og mange andre språk er liten grunn til å skille mellom såkalt indirekte og direkte objekt, og at man i stedet kan snakke om doble objekter. I norsk grammatikk har man riktignok så vidt meg bekjent tradisjonelt holdt seg til den klassiske analysen med recipient som indirekte objekt og theme som direkte objekt, men jeg velger i denne oppgaven å bruke analysen ovenfor også for norsk, uten at dette valget er spesielt viktig for den videre undersøkelsen. Det må imidlertid nevnes at Lødrup (1995) argumenterer for at det er den tradisjonelle analysen som er den korrekte for norsk, selv om han ikke kan forklare alle forhold rundt passivering.)

I henhold til Lexical Mapping Theory (LMT) i LFG vil de to alternativene ovenfor oppstå som resultat av følgende analyser (tilpasset etter Lødrup (2011)), der grammatisk funksjon er dekomponert og utledes av hvilke trekk (± restricted, ± object) LMT spesifiserer for rollene i en gitt argumentstruktur. De to ulike alternativene muliggjøres av at både recipient og theme er såkalt patiens-liknende roller, og bygger på at recipienten kan behandles på to alternative måter. Dekomponeringen av de fire fundamentale grammatiske funsjonene skjer i henhold til følgende matrise:

| | [-restricted} | [+restricted] |
|---|---|---|
| [-object] | SUBJ | $OBL_{\theta}$ |
| [+object] | OBJ | $OBJ_{\theta}$ |

Dobbelt-objekt-konstruksjonen:

give < agent recipient theme>

| | | | |
|---|---|---|---|
| [-o] | [-r] | [+o] | patiens-liknende (recipient) er [-restricted], sekundær patiens-liknende (theme) er [+object], andre (agent) er [-object] |
| \| | [+o] | [+r] | default: legg til + for uspesifiserte trekk |
| SUBJ | OBJ | $OBJ_{THEME}$ | |

Preposisjonsfrase-konstruksjonen:

| | | | | |
|---|---|---|---|---|
| give < | agent | recipient | theme> | |
| | [-o] | [-o] | [-r] | patiens-liknende (theme) er [-restricted], andre (agent og recipient) er [-object] |
| | \| | [+r] | [+o] | default: legg til + for uspesifiserte trekk |
| | SUBJ | $OBL_{RECIPIENT}$ | OBJ | |

Forklaringer som hevder at alternativene har ulik betydning tilbakefører alterneringen til mappingen fra argumentstruktur til syntaks, og opererer med to ulike såkalte event-strukturer for verb som kan alternere (Levin og Rappaport Hovav, 2005, s. 206-207): den ene strukturen viser til endring av lokasjon for theme-argumentet (realisert som PP: 'x cause y to be at z'), den andre til endring av possessor for theme-argumentet (realisert som dobbelt objekt: 'x cause z to have y'). Krifka (2003) oppgir blant andre Pinker (1989) og Speas (1990) som representanter for slike polysemi-baserte analyser. Pinkers analyse kan i følge Krifka angis som

- Dobbeltobjekt: [$_{event}$ give[Ann Beth[$_{state}$ HAVE Beth the car]]]
- Preposisjonsfrase: [$_{event}$ give[Ann the car[$_{event}$ GO the car[$_{path}$ to[$_{place}$Beth]]]]]

Speas (1990, s. 87) gir også to ulike leksikalske strukturer for *give*:

- GIVE y TO z: x cause y to come to be at (possession) z
- GIVE z y: x cause [z to come to be in STATE (of possession)] by means of [x cause [y to come to be at (poss) z]]

En fordel med denne siste analysen er at alternativene framstår klart som tilnærmede semantiske parafraser, ved at den første betydningen er inneholdt i den andre. Det kan også forklare hvorfor de ofte har identiske sannhetsbetingelser.

Forklaringer som legger avgjørende vekt på informasjonsstruktur ser bort fra slike finkornete forskjeller i betydning mellom alternativene. I stedet fokuseres slike faktorer som hvor "tunge" (lange eller komplekse) argumentene er relativt til hverandre, eller hvorvidt de representer ny (ukjent) informasjon eller ikke. I begge tilfeller (tungt eller nytt) vil argumentet tendere mot å flytte til høyre (sist) i setningen. Det avgjørende for valg av alternativ blir

dermed forholdet mellom informasjonsstatusen til recipient (eller benefactive) og theme (Levin og Rappaport Hovav, 2005, s. 217).

Krifka (2003, s. 3-6) går også gjennom et sett av mulige leksikalske restriksjoner for dativalterneringen i engelsk. Han hevder at dobbeltobjekt-konstruksjonen innebærer at recipienten direkte eller indirekte kommer til å possessere theme, slik at for eksempel

(2.3) ??Ann sent London a package

bare kan godtas hvis 'London' skal oppfattes som et metonym for en organisasjon. Videre hevder Krifka at preposisjonsfrase-konstruksjonen innebærer forflytning av theme, slik at for eksempel

(2.4) ??The explosion gave a headache to Beth

ikke kan godtas siden theme her bare oppstår i recipienten, ikke flyttes. Krifka hevder også at dobbeltobjekt-konstruksjonen krever at verbet ikke uttrykker kontinuerlig påvirkning av en kraft, slik at for eksempel

(2.5) ??Ann pulled Beth the box

ikke oppfyller denne restriksjonen, mens derimot

(2.6) Ann threw Beth the box

skulle være akseptabel. Legg her merke til at

(2.7) *Anne kastet Berit boksen

ikke går på norsk, slik at norsk må ha ytterligere restriksjoner på slike verb. I denne sammenheng er det interessant at Barðdal et al (2011, s. 60) nevner at verb som denoterer ballistisk bevegelse ikke kan forekomme i en dobbelt-objekt-konstruksjon i noe standard nord-germansk språk, til forskjell fra engelsk.

Når det gjelder kommunikasjonsverb i engelsk hevder Krifka (2003) at verb som angir måten noe sies på ikke tillater dobbeltobjekt, slik at for eksempel

(2.8) ??Ann shouted Beth the news

ikke går. Krifka hevder også at verb som angir forhindring av possesjon foretrekker dobbeltobjekt-konstruksjonen, slik at for eksempel

(2.9) ??Ann denied the icecream to Beth

ikke går. Til sist nevner Krifka en tendens til at bruk av dobbeltobjekt ofte innebærer at noe er fullført eller oppnådd, men at det samme ikke nødvendigvis trenger å være tilfelle ved bruk av preposisjonsfrase. Eksempelet her er forskjellen mellom (2.10) og (2.11):

(2.10) Beth taught the students French

(2.11) Beth taught French to the students

I følge Krifka kan studentene antas å ha lært fransk i det første tilfellet, men ikke nødvendigvis i det andre.

Barðdal et al (2011, s. 68-69) hevder at dobbelt-objekt-konstruksjonen i norsk kan brukes for følgende semantiske kategorier:

1. Faktisk overføring (gi noen noe, låne noen noe, betale noen noe, sende noen noe, bringe noen noe, skaffe noen noe)
2. Intensjon (love noen noe)
3. Skapelse (brøyte seg vei, koke seg noe)
4. Kommunikasjonsmåte (forklare noen noe, maile noen noe)
5. Muliggjøring (gjøre noen en tjeneste, nyttigjøre seg noe)
6. Tilbakeholding (forby noen noe, pålegge noen noe)
7. Mentale prosesser (forestille seg noe)

Barðdal et al går ikke inn på hvilke av disse som også tillater dativalternering.

## 2.1 Fra representasjon til valg av alternativ

Betydningsbaserte tilnærminger til ulike former for syntaktisk alternering har i hovedsak vært opptatt av hvordan alterneringen skal representeres innen sitt foretrukne syntaktiske rammeverk, og i mindre grad av hvilke faktorer som avgjør valg av alternativ i et konkret tilfelle. Dette har forandret seg med nyere tilnærminger basert på hensyn til informasjonsstruktur. Disse tar ofte utgangspunkt i empiriske data i form av et korpus, og ser på effekten av en, to eller flere utvalgte faktorer, og deres bidrag til alterneringen.

Arnold et al (2000, s. 35-39) viser for eksempel hvordan fraselengde og informasjonsstatus korrelerer med valg av alternativ for dativalternering (V NP NP kontra V NP PP) og 'heavy NP shift' (V NP PP kontra V PP NP) i engelsk. For begge alterneringene finner de at både relativ fraselengde (målt som differanse i antall ord mellom theme og recipient) og informasjonsstatus (vurdert som ny kontra gitt informasjon) korrelerer signifikant med alterneringen, slik at effekten ikke kan tilbakeføres til bare en av faktorene. For begge alterneringene finner de at relativt nye og tunge konstituenter tenderer til å komme sent i setningen, slik at for eksempel dativalterneringen tenderer til å manifestere seg som en dobbeltobjekt-konstruksjon hvis theme er relativt nyere og/eller tyngre enn recipienten. Undersøkelsen benytter et korpus med både talt og skrevet materiale (transkripsjoner av debatter i det kanadiske parlamentet), og benytter logistisk regresjon for å måle korrelasjonene. For dativalterneringen begrenser de seg til verbet *give*, fordi de mener at de for dette verbet kan se bort fra subtile semantiske forskjeller mellom alternativene.

Rosenbach (2007, s. 156-163) viser at animathet bidrar til valg av engelsk genitiv-variant (s-genitiv eller *of*-genitiv). Med referanse til flere andre korpus-baserte studier påviser hun at possessors animathet korrelerer med valg av alternativ uavhengig av andre faktorer som topikalitet og fraselengde. For å komme fram til dette har hun benyttet en metode med å holde alle andre faktorer konstant, i motsetning til ved logistisk regresjon, der man kan undersøke flere faktorer samtidig. Uansett finner hun at animat possessor tenderer til bruk av s-genitiv (*the girl's eyes*), mens ikke-animat possessor tenderer mot bruk av *of*-genitiv (*the roof of the house*), og forklarer dette med at prominens på animathet-skalaen tenderer mot å bruke den mer prominente prenominale spesifikator-posisjonen i s-genitiven enn den postnominale *of*-genitiven.

Aissen (2003, s. 450-470) viser hvordan animathet og/eller bestemthet bidrar til valg av kasus for såkalt 'differential object marking' i noen språk. I disse språkene trenger ikke objektet ha obligatorisk kasus, men kan alternere mellom eksplisitt kasus-markering, opsjonell kasus-markering eller ingen kasus-markering, avhengig av blant annet faktorene animathet og bestemthet. Hun undersøker blant andre hebraisk, tyrkisk, middelalder-spansk, hindi og persisk, og finner i alle tilfeller at høyere prominens på animathet- og/eller bestemthetskalaen samvarierer med mer bruk av obligatorisk kasusmarkering for objektet. I motsetning til de to ovennevnte studiene, er Aissens artikkel ikke basert på bruk av elektroniske korpus.

Rappaport Hovav og Levin (2008) argumenterer for at hvert enkelt verb har egne preferanser for dativalternering i engelsk. Eksempelvis nevner de at *give* oftest forkommer i en dobbeltobjekt-konstruksjon, mens *sell* oftest bruker preposisjonsfrase. De identifiserer tre brede klasser av dativ-verb: *give*-typen som signalerer at noe konkret gis til en mottaker, og som de hevder bare innebærer 'caused possession', *throw*-typen som signalerer øyeblikkelig forårsaking av ballistisk bevegelse, og *send*-typen som signalerer at noe sendes til en mottaker. De to siste typene hevder de innebærer både 'caused motion' og 'caused possession'. De hevder videre at *give*-typen av verb ikke koder for noen vei (path) til mottakeren, slik at det faktum at disse verbene også framviser dativalternering, ikke betyr at de to variantene innebærer noen semantisk forskjell på hvilken type hendelse (event) de representerer. For disse verbene fører de derfor alterneringen i hovedsak tilbake til hensyn som informasjonsstruktur og fraselengde, slik at gitt materiale kommer før nytt materiale, og tungt materiale kommer til slutt i setningen. De antyder at for eksempel en egen effekt av animathet egentlig er en konsekvens av informasjonsstruktur (s. 157), men skriver også (s. 159) at deres foreslåtte forklaring må underbygges av korpus-studier.

Størst bidrag til empirisk baserte undersøkelser av syntaktisk alternering har likevel Joan Bresnan og ulike medforfattere gitt i en serie artikler fra 2003 og senere, som alle behandler dativalternering i ulike varianter av engelsk. I disse artiklene påvises det at dativalternering avhenger av samspillet mellom en rekke faktorer, inklusive alle de ovennevnte. I disse artiklene er det en utvikling fra en kvalitativ analyse bygget på optimalitetsteori til en kvantitativ, probabilistisk analyse bygget på matematisk modellering. Slike matematiske modeller er formler som gjør det mulig å forutsi valg av alternativ ut fra et gitt sett av faktorer (parametre), med høy grad av treffsikkerhet. Slike modeller er rent deskriptive, og kan i seg selv ikke si noe om hvorfor faktorene bidrar i den retning de gjør. Forklaringen må i stedet

søkes i en overgripende teori om for eksempel prominens: høy prominens på en gitt skala for en bestemt faktor reflekteres i en prominent syntaktisk plassering av tilsvarende konstituent. 'Harmonic alignment' mellom prominensskala og syntaktisk posisjon kan være en slik teori. Man tar da utgangspunkt i antatt universelle prominensskalaer, som kan forenkles til binære sådanne, slik at høy prominens på for eksempel en aksesserbarhetsskala er assosiert med at konstituenten realiseres rett etter verbet, mens lav prominens på samme skala er assosiert med at konstituenten realiseres til slutt i setningen. Dermed vil en høy-prominent recipient forventes realisert som første objekt i en dobbelt-objekt-konstruksjon V-NP-NP, mens en lav-prominent recipient vil forventes realisert i en preposisjonsfrase (V-NP-PP). Tilsvarende vil et høy-prominent theme forventes realisert som det direkte objektet i en preposisjonsfrase-konstruksjon V-NP-PP, mens et lav-prominent theme vil forventes realisert som andre objekt i en dobbelt-objekt-konstruksjon V-NP-NP. Siden man må forvente å bruke flere slike binære prominensskalaer samtidig, og de til dels kan gi motstridende preferanser, blir det viktig å kombinere dette teoretiske utgangspunktet med komputasjonelle teknikker som tillater en å formulere det kategoriske valget mellom alternativene i form av et probabilistisk samspill mellom de faktorene og prominensskalaene som inngår i denne formen for 'harmonic alignment'. Matematisk modellering er en slik teknikk.

Dette arbeidet innledes i Bresnan og Nikitina (2003) med å se på problemet innenfor rammen av såkalt stokastisk optimalitetsteori. I denne artikkelen viser de hvordan noen av faktorene kan realiseres som skranker (constraints), men kvantifiserer ikke det relative bidraget hver skranke gir til valg av konstruksjonstype for recipienten. De hevder innledningsvis at det tidligere har blitt framsatt to ulike typer forklaringer på hva som driver dativalterneringen. På den ene siden finnes semantiske tilnærminger som underkjenner at det foreligger en ekte alternering fordi de opererer med et antall semantiske klasser av dativverb (for eksempel verb som innebærer påvirkning av en kraft, kommunikasjonsverb, etc) og idiomer (*give someone a headache*, *give someone a punch*) som hver for seg har en unik syntaks bestemt av den underliggende betydningen. På den andre siden finnes en type tilnærminger som tar utgangspunkt i kontekstuelle faktorer (i bred forstand) som informasjonsstruktur, animathet, bestemthet og den relative kompleksiteten til de to objektene. I artikkelen går Bresnan og Nikitina gjennom flere påståtte semantiske klasser og påviser ved korpussøk at disse ikke entydig bestemmer valget av alternativ, og at dativalterneringen forekommer ved mange flere verb og idiomer enn man tidligere har trodd. På denne bakgrunn avviser de en rent semantisk tilnærming, og legger i stedet fram en alternativ modell innenfor rammen av stokastisk

optimalitetsteori, som nevnt ovenfor. (Krifka (2003), som selv argumenterer for et polysemi-basert syn, men også tar høyde for at informasjonsstruktur kan avgjøre valget i en gitt kontekst, påpeker at Bresnan og Nikitina dermed argumenterer for informasjonsstrukturens betydning for valg av alternativ ut fra et monosemi-basert syn på dativalternering.)

I Bresnan et al (2007) legges det fram en rent probabilistisk modell for dette valget, der hver faktors bidrag er kvantifisert og bestemt ut fra et talespråkskorpus (SWITCHBOARD-korpuset for telefonsamtaler) ved hjelp av multifaktoriell logistisk regresjon. I artikkelen gjennomgås et antall argumenter mot å bruke korpusdata på denne måten, som alle avvises på statistisk grunnlag.

For det første viser de at det at faktorene til dels er korrelerte med hverandre, ikke betyr at de kan reduseres til bare en faktor, som for eksempel syntaktisk kompleksitet. Dette gjøres ved å lage en eksplisitt matematisk (probabilistisk) modell for valg av dativalternativ, modell A, basert på korpuset, der hver faktor kan vises å være individuelt signifikant for valget. Siden jeg i denne oppgaven i hovedsak vil replikere dette for norske data, viser jeg til senere kapitler for detaljer.

For det andre viser Bresnan et al at det ikke er noe problem at korpuset pooler data fra mange individer.  Dataene i SWITCHBOARD-korpuset ble kodet for taler, noe som gjorde det mulig å kontrollere for individuelle preferanser. Gjennom en form for randomisert resampling av individuelle data, kunne de remodellere og sjekke om individuelle forskjeller overstyrte responsene i den grunnleggende modellen (med poolede data fra alle talerne). Det viste seg å ikke være tilfelle. Fellestrekkene mellom talerne hva valg av dativalternativ angår, var langt viktigere enn de individuelle forskjellene.

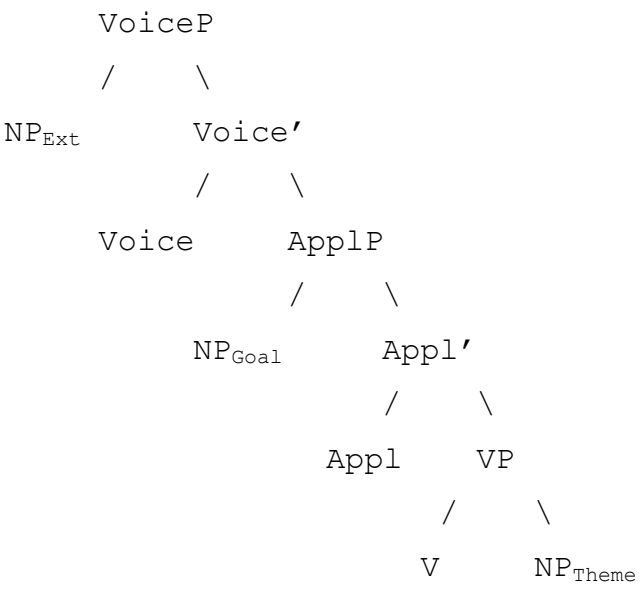
For det tredje viser Bresnan et al at det heller ikke er noe problem å poole data fra ulike verb, selv om hvert enkelt verb skulle ha egne preferanser for dativalternativ. De brukte en alternativ modelleringsteknikk kalt multinivå logistisk regresjon eller mixed-effect logistisk regresjon, som gjør det mulig å ta hensyn til hvert enkelt verbs ulike bruksmåter. På denne måten kunne de kode for inntil fem ulike bruksmåter fordelt på i alt 38 verb, som totalt ga 55 ulike betydninger, hvorav abstrakt bruk av *give* alene stod for en tredjedel av dataene. Den nye modellen, modell B, viser at faktorene fra modell A fortsatt er individuelt signifikante, og virker i samme retning hva valg av dativalternativ angår.

For det fjerde viser Bresnan et al at det ikke er noe problem med forskjeller mellom korpus. De laget en lett revidert modell, modell C, på basis av både SWITCHBOARD-korpuset og Wall Street Journal, og viser at denne modellen har en probabilistisk struktur som gjør at den er like godt tilpasset taledata som skriftdata, til tross for til dels store forskjeller mellom del-korpusene i bruk av dativalternativ.

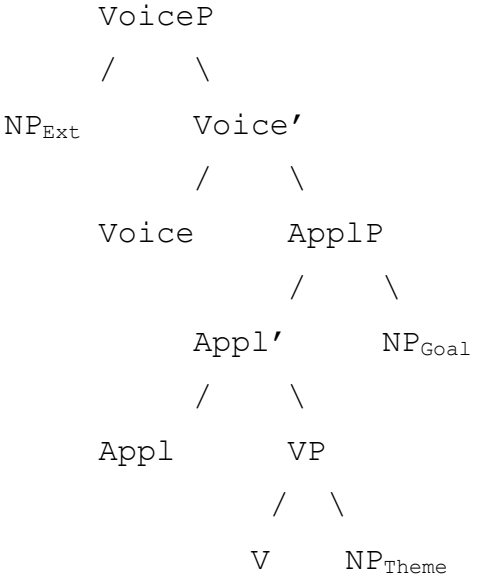
For å oppsummere: Bresnan et al (2007) viser at reduktive teorier som tilbakefører alterneringen til bare en av faktorene ikke kan redegjøre for alle fakta, at det faktum at man bruker data fra mange ulike språkbrukere ikke undergraver modellenes gyldighet, at valg av alternativ ikke i sin helhet avgjøres av det enkelte verbs leksikalske bias, og at forskjeller mellom korpus ikke umuliggjør å bruke en gitt modell på ukjente data. Modellen(e) som framsettes tar utgangspunkt i faktorene semantisk klasse, hvorvidt recipient og theme er kontekstuelt gitt, hvorvidt de er pronomener, hvorvidt de er definitte, hvorvidt de er animate, person, tall, og relativ lengde. Modellen(e) viser seg å kunne predikere valg av alternativ med inntil 94 % treffsikkerhet. Artikkelen behandler ikke eksplisitt spørsmålet om hvorvidt modellen har psykologisk relevanse, og heller ikke om den i så fall angår hva Chomsky kalte språklig kompetanse eller performanse, men man må i alle fall kunne si at den har teoretisk og praktisk interesse for blant annet anvendt lingvistikk (for eksempel andrespråkslæring) og datalingvistikk.

Et problem med denne modellen er at det ikke er gitt at den representerer en psykologisk realitet for den enkelte språkbruker, men i Bresnan (2007) og Bresnan og Ford (2010) legges det fram data fra ulike eksperimenter (spørreundersøkelser) som synes å bekrefte at slike probabilistiske modeller kan tilsvare intuisjonene til vanlige språkbrukere.

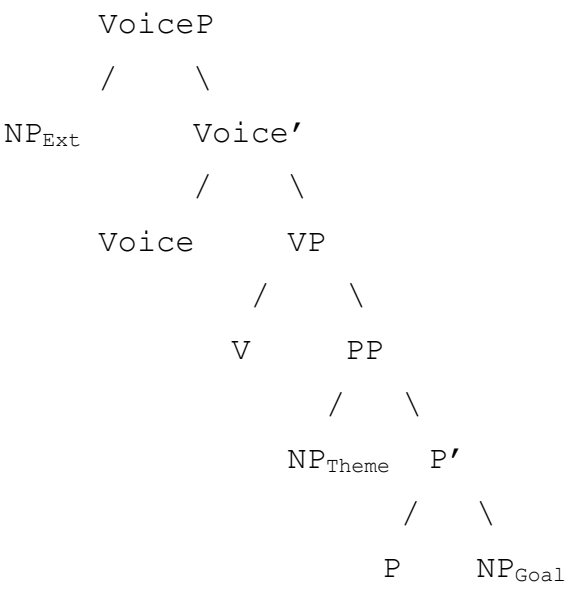
Bresnans arbeider i denne retning har også blitt kritisert, og i en artikkel som er et direkte tilsvar til Bresnan et als ulike artikler, forsvarer Bruening (2010) et struktur-drevet, polysemi-basert syn på dativalternering. Han argumenterer blant annet ved å se detaljert på kvantifikator-rekkevidde og andre tekniske hensyn som ikke har blitt spesielt vektlagt i den ovenfor nevnte litteraturen, med at noen tilsynelatende tilfeller av preposisjons-dativ-konstruksjoner egentlig er dobbelt-objekt-konstruksjoner som har undergått hva han kaller R-dativskift. Bruening (s. 289-90) antar følgende syntaktiske trær for henholdsvis dobbelt-objekt-konstruksjonen, en R-dativskiftet dobbelt-objekt-konstruksjon og preposisjons-dativ-konstruksjonen:




Dobbelt-objekt-konstruksjonen




R-dativskiftet dobbelt-objekt-konstruksjon



Preposisjonsdativ-konstruksjonen

En R-dativskiftet dobbelt-objekt-konstruksjon krever i følge Bruening at Goal (recipient) markeres med preposisjonen *to*, og setningen vil da ha samme linearisering som en preposisjonsdativ-konstruksjon. I følge Bruening kan denne R-dativskiftede konstruksjonstypen forklare hvorfor noen idiomer (*give someone a headache*) som man tidligere antok ikke kunne forekomme med preposisjonsdativ, tilsynelatende likevel gjør det, når man leter i svært store korpus. Han konkluderer derfor, kontra Bresnan et al, med at grammatikken skiller mellom dobbelt-objekt-konstruksjoner og preposisjons-dativ-konstruksjoner både syntaktisk og semantisk. I en fotnote avviser han også at en eventuell probabilistisk algoritme for valg av alternativ skulle være en del av språkbrukerens interne grammatikk, og hevder at den like gjerne kan være en del av vedkommendes språkbruk (performanse).

Colleman (2009) er også kritisk til Bresnan et als konklusjoner. Han mener at deres påstand om at semantisk baserte forklaringer på dativalternering ikke holder, ikke følger av det faktum at de i store korpus har funnet moteksempler til hva slike hypoteser forutsier om valg av alternativ. Vel kaster deres funn tvil over lingvisters introspeksjon om hva som er grammatiske setninger, og spesielt om hva som *ikke* er grammatiske setninger, men like fullt kan ulike verbs statistiske preferanser for valgt alternativ brukes som leksikalsk støtte i diskusjoner om det semantiske forholdet mellom de to konstruksjonene. Collemans artikkel

viser en alternativ bruk av korpus i forhold til Bresnan et al. Mens Bresnan et al bruker korpuset som datagrunnlag for å bygge en probabilistisk modell for *valg* av alternativ (i hovedsak med utgangspunkt i informasjonsstruktur og lignende faktorer), bruker Colleman korpuset som datagrunnlag for å *klassifisere* de enkelte ditransitive verbs preferanser for alternativ, og *kvantifisere* hvor sterk preferansen er. Med bakgrunn i denne klassifiseringen kan han komme med semantiske generaliseringer om preferansene til ulike klasser av verb, og påvise eventuelle avvik fra tidligere klassifiseringer som i hovedsak var grunnet på den enkelte lingvists introspeksjon. Mens engelsk og nederlandsk på makronivå har felles generaliseringer, påviser han at på mikronivå, det enkelte verb og dets nærmeste oversettelse i det andre språket, kan det være forskjeller i preferanse. (Jamfør min kommentar om *throw* og *kaste* ovenfor.)

Newmeyer (2002) representerer en prinsipiell kritikk av stokastisk optimalitetsteori fra et generativt ståsted. Newmeyer er spesielt kritisk til at UG skal inneholde tematiske og relasjonelle hierarkier. Han hevder at grammatiske prosesser eventuelt kan referere til individuelle tematiske roller, ikke et helt hierarki, og at grammatiske relasjoner ikke er medfødte (’innate’) kategorier. Bresnan og Aissen (2002) er et tilsvar til Newmeyer, der hans argumenter tilbakevises punkt for punkt. De konkluderer (s. 92) med at selve konseptet ’grammatikk’ har blitt endret, ved at generering av strukturer (trær) ikke lenger er spesielt viktig, sammenliknet med å finne og funksjonelt motivere skranker (constraints).

## 2.2 Det spesielle verbet ’gi’

Mange forfattere kommenterer at verbet ’gi’ er spesielt. Newman (1998) skriver i forordet til en bok angående lingvistikken rundt verbet ’gi’, at dette verbet er semantisk komplekst og samtidig grunnleggende for hvordan vi erfarer verden. Han hevder for eksempel at det er et av de første verbene et barn forstår, spesielt i form av ’give me …’-konstruksjoner. Den semantiske kompleksiteten består i følge Newman i at:

- Der er tre viktige entiteter (en giver, en ting som overføres, og en mottaker)
- Der er en interaksjon mellom giveren og tingen

- Der er en interaksjon mellom mottakeren og tingen
- Der er bevegelse av tingen fra giver til mottaker
- Der er forandring i kontroll over tingen, som går fra giver til mottaker
- I den mest typiske formen for giving, er hendene til både giver og mottaker involvert
- Givingen er gjort intensjonelt
- Givingen er vanligvis gjort til fordel for mottakeren slik at mottakeren kan gjøre bruk av den overførte tingen

Som en ser av denne listen gjelder dette giving som fysisk overføring av råderett over en ting, men Newman nevner også at den semantiske kompleksiteten gjør at verbet ofte brukes i metaforisk og figurlig betydning. Newman anser videre at den store variasjonen i syntaktisk realisering av denne typen verb i verdens språk kommer av at denne handlingen er så erfaringsmessig grunnleggende og semantisk kompleks.

Kittilä (2006) hevder i en artikkel som vel og merke bare tar opp giving som fysisk overføring av ting, at ’gi’ er et atypisk treverdig verb på mange måter. Han hevder for eksempel at ’gi’ krever en eksplisitt referanse til alle tre deltakerne ’rather mandatorily’ (s. 585), så vel i engelsk som i for eksempel norsk. En mulig forklaring finner han i at ’gi’ er et ganske uinformativt verb som bare referer til at en ting overføres fra en giver til en mottaker, uten ytterligere informasjon. På basis av sin tverrlingvistiske studie formulerer han noen universalier, hvorav den andre (s. 605) er spesielt interessant her:

- Universalie 2: Hvis et språk har en morfosyntaktisk operasjon som dativskift, som opsjonelt promoverer recipienten til direkte objekt, så er ’gi’ i klassen av verb som tillater dette.

Rappaport Hovav og Levin (2008) er helt eksplisitte (s. 133) på at valg av dativalternering i engelsk er bestemt av informasjonsstruktur og frasekompleksitet hva angår verb av typen ’give’. Når det gjelder det spesielle verbet *give* hevder det at dette leksikaliserer ’caused possession’ og ingenting annet (s. 135), og at dets rotbetydning ikke tilfører noe utover det som allerede ligger i ’caused possession’ som eventskjema.

Verbet *gi* synes derfor å være spesielt velegnet for en multifaktoriell undersøkelse av hvilke faktorer som påvirker valg av dativalternativ, og er hovedfokus for denne oppgaven.

## 2.3 Teknikker brukt for å håndtere flere faktorer

Aissen (2003, s. 440-449) bruker i sin artikkel om såkalt ’differential object marking’ en teknikk fra optimalitetsteori, kalt ’harmonic alignment’, for blant annet å håndtere samspillet mellom animathet og definitthet. ’Harmonic alignment’ tar utgangspunkt i to separate prominens-skalaer, hvorav en må være binær, og produserer subhierarkier av skranker som uttrykker hvor markert hver enkelt mulig assosiasjon mellom elementene på disse to skalaene er. I Aissens tilfelle er disse skalaene:

- Relasjon: subjekt > objekt, og
- Animathet: human > animate > inanimate, eller
- Definitthet: pronoun > name > definite > indefinite specific > nonspecific

‘Harmonic alignment’ mellom for eksempel relasjonsskalaen og animathet gir opphav til tilsvarende subhierarkier mellom skranker, i dette tilfellet

- *OBJ/HUM >> *OBJ/ANIM >> *OBJ/INAN
- *SUBJ/INAN >> *SUBJ/ANIM >> *SUBJ/HUM

Disse subhierarkiene konjungeres lokalt med en skranke $*0_C$ som motvirker *fravær* av eksplisitt kasusmarkering. Til slutt interpoleres en skranke $*STRUC_C$ som motvirker *bruk* av eksplisitt kasusmarkering slik at det resulterende subhierarkiet reflekterer den faktiske bruken av kasusmarkering i et gitt språk. Aissen hevder (s. 458) for eksempel at i språket Dhargari er alle animate objekter kasus-markert, noe som oppnås ved subhierarkiet

$$*OBJ/HUM \;\&\; *0_C \;>>\; *OBJ/ANIM \;\&\; *0_C >> *STRUC_C >> *OBJ/INAN \;\&\; *0_C$$

På denne måten får Aissen et teknisk verktøy til å undersøke hvordan valg av kasus ved ’differential object marking’ avhenger av grad av markerthet. Problemet med denne teknikken er at den vanskelig kan utvides til å behandle flere faktorer samtidig. Aissen har løst problemet for to faktorer ved å ta kryssproduktet av skalaene, og dele den resulterende partielle rangeringen av relativ markerthet i tre soner for henholdsvis obligatorisk

kasusmarkering, opsjonell kasusmarkering og ingen kasusmarkering. For at sonen for opsjonell kasusmarkering skal gi korrekte resultater foreslår hun å benytte stokastisk optimalitetsteori under evalueringen for å tillate dynamisk rerangering av skranker. Aissen sier ikke noe om hvordan teknikkene hennes kan utvides til tre eller flere faktorer. Det er klart at en tilsvarende partiell rangering da ville bli svært komplisert å håndtere.

Bresnan og Nikitina (2003) bruker i sin første artikkel om dativalternering også stokastisk optimalitetsteori for å håndtere samspillet mellom flere skranker. I stokastisk optimalitetsteori er skrankene ikke bare rangert relativt til hverandre, men hver skranke er tilordnet en normalfordelt sannsynlighetsfunksjon for hvor skranken faller på et kontinuum av mulige rangeringspunkter. Dette muliggjør at skrankene i et konkret tilfelle er rerangert slik at utfallet av evalueringen blir forskjellig fra normalen. Uten at de går i detaljer bruker også Bresnan og Nikitina en teknikk som gjør det mulig å håndtere flere faktorer samtidig, men siden denne bygger på den samme typen 'harmonic alignment' mellom to og to skalaer som Aissen bruker, blir den komplisert å håndtere når mange faktorer inkluderes. I artikkelen ser de også på hvordan leksikalsk variasjon kan introduseres ved hjelp av skranker som er spesifikke for enkeltverb eller finkornete klasser av verb, noe som vel til en viss grad strider mot ideen bak optimalitetsteori.

Bresnan et al (2007) introduserer bruk av logistisk regresjon for å modellere valg av dativalternativ. Dette er en probabilistisk teknikk som på en forholdsvis enkel måte gjør det mulig å vise samspillet mellom mange ulike faktorer uten å måtte sette opp et stort teoretisk og teknisk apparat for å få det til. På den andre siden er man helt avhengig av bruk av IT for å gjennomføre modelleringen. De resulterende matematiske modellene har den fordel at de følger samme prinsipper som tilsvarende modeller innen mange andre fagområder innen humaniora, samfunnsvitenskap og naturvitenskap, og følgelig vil være tverrvitenskapelig forståelige. Siden dette også er den metoden jeg vil bruke, viser jeg til neste kapittel for detaljer.

# 3 Teori og metode

Denne oppgaven bruker såkalt logistisk regresjon (se for eksempel Dalgaard (2008, kapittel 13)) for å beskrive og forutsi dativalternering (V NP NP eller V NP PP) i norsk ut fra data hentet fra et korpus. Logistisk regresjon, også kalt logistisk modellering, er en av flere ulike såkalte generaliserte lineære modeller (se også Manning (2003, s. 332-336) for en introduksjon), som er en familie av matematiske modeller som kan brukes til å beskrive fenomener som avhenger av et antall ulike faktorer, også kalt parametre. Logistisk modellering kan brukes for å beskrive sannsynligheten for en bestemt hendelse (i dette tilfellet at dativalterneringen er realisert som V NP PP) gitt et antall kategoriske og/eller numeriske parametre og parameterkoeffisienter (også kalt modellkoeffisienter eller regresjons-koeffisienter) som kan beregnes ut fra et stort datasett (sample). Datasettet bør inneholde flere hundre enkelthendelser (eventer) for at koeffisientene skal beregnes med tilstrekkelig grad av signifikans, og antall positive hendelser (her: V NP PP i stedet for V NP NP) bør være minst ti ganger så stort som antall parametre som modellen tar hensyn til.

Modellen kalles logistisk fordi den bygger på den såkalte logistiske funksjonen $f(z)=1/(1+e^{-z})$, som tar vilkårlig store negative og positive argumenter, og som funksjonsverdi gir et tall mellom 0 og 1, som dermed kan tolkes som en sannsynlighet. Variabelen z er en såkalt lineær kombinasjon av de faktiske parameterverdiene $x_i$ vektet med de beregnede modell-koeffisientene $\beta_i$. En positiv modellkoeffisient angir at den tilsvarende parameteren bidrar til at hendelsen (her: V NP PP) inntreffer, mens en negativ koeffisient angir at den tilsvarende parameteren bidrar til at hendelsen *ikke* inntreffer (altså her: V NP NP i stedet). Absoluttverdien av en koeffisient sier noe om i hvor stor grad den tilsvarende parameteren bidrar i den ene eller andre retningen, slik at koeffisientverdier nær null angir at parameteren i liten grad er virksom i modellen. Beregningen av koeffisientene gjøres ved et såkalt 'maximum likelihood' estimat basert på det underliggende datasettet, og kan gjøres automatisk som en del av modelleringen i statistikkprogrammet R som jeg bruker.

Det er viktig å være klar over at en slik matematisk modell ikke sier noe om *hvorfor* en bestemt kombinasjon av faktorer tenderer til å gi opphav til en bestemt hendelse. Modellen kan heller ikke brukes til å forklare hvorfor en bestemt faktor bidrar i en bestemt retning. Modellen kan bare vise hvilke sammenhenger og mønstre det underliggende datamaterialet

oppviser, og kvantifisere hver enkelt faktors bidrag til en bestemt type hendelse. Videre er det nødvendig å bruke modelleksterne kriterier for valg av hvilke faktorer modellen skal ta hensyn til. Dette valget avhenger altså i dette tilfellet av hva vi på forhånd mener å vite om hvilke pragmatiske, semantiske og formelle faktorer som påvirker dativalternering i norsk. Fra engelsk vet man (se spesielt Bresnan et als mange artikler) at animathet, definitthet, pronominalitet, aksesserbarhet og fraselengde av recipient og theme kan påvirke valg av alternativ. Det enkelte verb og dets spesifikke semantiske betydning kan også virke inn. Videre kan modus (skrift eller tale) bidra til valget. Til sist kan det nevnes at noen andre faktorer også har vært nevnt (person og tall for recipient og theme; priming ved at en bestemt konstruksjon nettopp har vært brukt). Med unntak av disse siste har jeg valgt å se på de samme faktorene i norsk.

Når det gjelder det praktiske arbeidet burde ideelt sett korpuset man bruker være tagget syntaktisk slik at man kunne hente ut alle forekomster av indirekte objekt (realisert som sådan eller som objekt til preposisjonen *til*). Dermed kunne man få med alle ditransitive verb i sin rent ditransitive bruk. I realiteten finnes det ikke noe slikt korpus tilgjengelig, slik at jeg i stedet har måttet bruke morfologisk og leksikografisk taggede korpus til å hente ut datamaterialet. Metoden har da vært å søke etter alle forekomster av et gitt verb, og så sile bort all ikke-ditransitiv bruk av verbene manuelt. Resultatet er en tekstfil bestående av relevante treff fra korpuset med en minimal kontekst for hvert treff. Hvert treff har så blitt kodet i henhold til følgende regler:

1. Realisering: NP betyr dobbeltobjekt-konstruksjon, mens PP betyr preposisjonsfrase.
2. Verb: infinitivsformen av verbet.
3. Bruk: a betyr abstrakt bruk, c betyr kommunikativ bruk, t betyr fysisk overføring.
4. Verbbruk: Verb.Bruk
5. AnimRec: animat recipient kodes som 1, ellers 0.
6. AnimTh: animat theme kodes som 1, ellers 0.
7. DefRec: definitt recipient kodes som 1, ellers 0.
8. DefTh: definitt theme kodes som 1, ellers 0.
9. ProRec: pronominal recipient kodes som 1, ellers 0.
10. ProTh: pronominal theme kodes som 1, ellers 0.
11. AccRec: aksesserbar recipient kodes som 1, ellers 0.

12. AccTh: aksesserbar theme kodes som 1, ellers 0.
13. LRec: fraselengde av recipient i antall ord.
14. LTh: fraselengde av theme i antall ord.
15. Kilde: Bokstavkoder som identifiserer delkorpuset treffet stammer fra.

Animathet, definitthet, pronominalitet og aksesserbarhet baserer seg i noen grad på tradisjonelle skalaer eller hierarkier. I stedet for å kode for alle punkter på skalaene, har jeg valgt et bruddpunkt for hver skala slik at alle punkter til høyre for dette bruddpunktet kodes som 0, og resten (de mest prominente) som 1. Dette er gjort delvis av praktiske grunner (lettere koding: mindre å skrive), delvis av metodiske (noen skalapunkter er så dårlig representert at de ikke ville gi en brukbar modell).

Følgende kommentarer er på sin plass:

**Animathet**: I Newmeyer (2002, s. 49) gjengis følgende hierarki:

- 1st person pronoun > 2nd person pronoun > 3rd person pronoun > proper noun > human common noun > animate common noun > inanimate common noun.

I praksis har jeg brukt en versjon der bruddpunktet ligger inne i inanimate common noun slik at mennesker, dyr, og uorganiserte grupper av mennesker kodes som animate (1), og alt annet, inklusive organisasjoner, som ikke-animate (0).

**Definitthet**: Newmeyer (2002, s. 49) gjengir følgende hierarki:

- personal pronoun > proper noun > definite NP > indefinite specific NP > non-specific NP.

Jeg har valgt å legge bruddpunktet på definite NP, slik at *den rike mannen* kodes som definitt (1), mens *en rik mann* kodes som ikke-definitt (0).

**Pronominalitet**: Personlige, demonstrative og refleksive pronomener kodes som pronominale (1), dessuten også nomenfraser innledet med possessivpronomen, mens alt annet kodes som ikke-pronominal (0).

**Aksesserbarhet**: Hvis referenten er nevnt i umiddelbar foregående kontekst kodes den som aksesserbar (1), likeledes for 1. og 2. person pronomen, samt 3. person pronomen og egennavn hvis det ut fra sammenhengen er overveiende sannsynlig at disse var nevnt før den

umiddelbart foregående konteksten. Dette siste kan synes noe vagt, men er etter beste skjønn anvendt for å kompensere for at konteksten er relativt kort, typisk mindre enn 70 tegn.

**Relativ lengde**: For å unngå at ekstremverdier skal påvirke modellen for mye må relativ lengde lineariseres ved å ta logaritmen av differansen mellom fraselengdene for recipient og theme. Dette kan enkelt gjøres i statistikkverktøyet (programmet) R før modellering.

**Verbbruk**: Gjør det mulig å splitte hvert enkelt verb i inntil tre semantisk ulike bruksmåter. Dette er selvfølgelig bare nyttig hvis man ser på flere verb samtidig. Hvis man bare ser på ett verb vil parameteren Bruk alene gi samme effekt.

Etter koding og redigering har jeg en tekstfil (se Vedlegg for eksempler) der hvert brukbare treff forutgås av en linje på formen

(4.1) NP gi a gi.a 1 0 1 0 1 0 1 0 1 2 Af

som i dette tilfelle kunne representere det mulige treffet

(4.2) Hun så ham og ga ham en klem.

Dette eksempelet koder da for dobbeltobjekt, verbet *gi* i abstrakt betydning, animat/ definitt/ pronominal/ aksesserbar recipient, ikke-animat/ ikke-definitt/ ikke-pronominal/ ikke-aksesserbar theme, relativ fraselengde 1/2 (recipient/theme), kilde Aftenposten. Hver slik linje kan brukes som input til bestemte prosedyrer i programmet R. Den prosedyren jeg trenger heter **glm** (generalized linear model) og kan blant annet brukes til å lage multivariable logistiske regresjonsmodeller av den typen jeg trenger. Framgangsmåten er beskrevet i kapittelet om dativalternering i engelsk i Johnson (2008: kapittel 7.4), og tilpasses her.

Norsk Referansegrammatikk (2006, kapittel 8.4.4) oppgir følgende verb som ditransitive (inndelingen er som i boka):

1. Verb som veksler mellom nominal og preposisjonsfrase
    a. Overføring av gjenstander og lignende
        i. Bringe, by, gi, leie, låne, overlate, overrekke, rekke, selge, sende, servere, skaffe, skjenke, tilby, tildele, tilegne
    b. Pengesummer

        i. Betale, bevilge, løyve
    c. Meningsinnhold
        i. Betro, fortelle, henstille, love, meddele, si, forklare
    d. Sanseinntrykk
        i. Syne, vise
    e. Uklare
        i. Anbefale, anvise
2. Verb som bare tar nominal
    a. Overføring
        i. Befale, beordre, forsikre, innbille, innvilge, minne, oppfordre, oppmode, svare
    b. Nektelse
        i. Berøve, forby, koste, nekte
    c. Spesielle semantiske forhold
        i. Anmode, be, bebreide, forespeile, lære, skylde, underkaste, unne, misunne, unnskylde, volde, ønske
    d. Sammensatt med preposisjoner
        i. Fradømme, frakjenne, frata, tilføre, tilgi, tilkjenne, tillate, tillegge, tilrå, tilsette, tiltro, idømme, påføre, pålegge, påtvinge
    e. Refleksive
        i. Foreta seg, nærme seg, tilegne seg, tilrane seg, tiltvinge seg, underlegge seg

Av disse har jeg sett på de aller fleste verbene i gruppe 1. Selv om det hadde vært interessant å se om noen av verbene i gruppe 2 også kunne alternere i spesielle tilfeller, regnet jeg med at det i så fall ville skje så sjelden at disse verbene ikke burde inngå i modelleringen. Jeg har begrenset undersøkelsen til ditransitiver med nominal eller preposisjonsfrase med *til*, og sett bort fra preposisjonsfraser med *for*.

Jeg så til å begynne med på korpuset "Oslo-korpuset av taggede norske tekster", nærmere bestemt delkorpuset bestående av treff fra Aftenposten. Av praktiske grunner begrenset jeg alle søk i dette korpuset til finitte verbformer. Det viste seg at de fleste av de undersøkte verbene i all hovedsak ble brukt ikke-ditransitivt i korpuset, slik at det ble en uforholdsmessig stor jobb å skaffe et tilstrekkelig antall gode treff.

Følgende tabell viser fordelingen på treff og ditransitiv bruk for de undersøkte verbene (passiver og verbene *si* og *vise* er ikke analysert):

| Verb | Antall treff | Ditransitiv bruk | Prosent |
|---|---|---|---|
| bringe | 68 | 7 | 10,3 % |
| by | 58 | 5 | 8,6 % |
| gi | 1087 | 352 | 32,4 % |
| leie | 25 | 0 | 0,0 % |
| låne | 21 | 5 | 23,8 % |
| overlate | 21 | 9 | 42,9 % |
| overrekke | 7 | 4 | 57,1 % |
| rekke | 34 | 1 | 2,9 % |
| selge | 148 | 11 | 7,4 % |
| sende | 159 | 28 | 17,6 % |
| servere | 28 | 1 | 3,6 % |
| skaffe | 29 | 22 | 75,9 % |
| skjenke | 3 | 2 | 66,7 % |
| tilby | 69 | 14 | 20,3 % |
| tildele | 5 | 2 | 40,0 % |
| tilegne | 2 | 2 | 100,0 % |
| betale | 102 | 6 | 5,9 % |
| bevilge | 20 | 10 | 50,0 % |
| betro | 10 | 1 | 10,0 % |
| fortelle | 566 | 30 | 5,3 % |
| henstille | 0 | 0 | |
| love | 111 | 8 | 7,2 % |
| meddele | 4 | 0 | 0,0 % |
| si | 5097 | | |
| vise | 820 | | |

Tabell: Ditransitive verb i Aftenposten

På bakgrunn av disse tallene valgte jeg å gå over til bare å se på verbet *gi*, men bruke et større delkorpus bestående av alle kilder som var tagget som aviser. Jeg utelukket nå både infinitte former og passiver fra søket. Verbet *gi* brukes her ditransitivt i ca hvert tredje korpustreff (2727 av totalt 7624 treff), og forekommer dessuten hyppigst av alle verb som oppgis som ditransitive i Norsk Referansegrammatikk (med unntak av *si*, som jeg antok i hovedsak brukes på andre måter). For å sammenlikne skriftspråk og talespråk undersøkte jeg deretter det samme verbet *gi* i korpusene BigBrother-korpuset, NoTa-Oslo, og TAUS. Siden disse

korpusene bare tilbyr lemmasøk, kunne jeg her ikke utelukke passiver og infinitte former. Treffene fra disse tre korpusene ble samlet til én tekstfil for å få nok treff til en egen modellering av talespråk. Denne modellen bygger imidlertid på et så lite datagrunnlag (200 relevante av totalt 607 treff) at den ikke er særlig signifikant.

Om korpusene jeg har brukt skriver Tekstlaboratoriet (2012) følgende på sine web-sider:

- Oslo-korpuset av taggede norske tekster "er et norsk skriftspråkskorpus med 18,5 millioner ord for bokmål og 3,8 millioner ord for nynorsk. Tekstene er hentet fra aviser, ukeblad, romaner og offentlige dokument. Korpuset er tagget med Oslo-Bergen-taggeren."

- BigBrother-korpuset "er et talespråkskorpus som består av ortografiske transkripsjoner knyttet til lyd og videoopptak fraTVNorges Big Brother-sendinger fra 2001."

- NoTa-Oslo "er et talespråkskorpus med opptak fra 2004 - 2006 med ca 900000 ord knyttet til lyd- og videoopptak. Korpuset inneholder ortografiske transkripsjoner av samtaler og intervju fra informanter født og oppvokst i Oslo-området. 144 av informantene er representative med hensyn til alder, kjønn, bosted og utdannelse."

- TAUS "er et talespråkskorpus fra Oslo. Materialet til TAUS ble samlet inn ved Universitetet i Oslo tidlig på 1970-tallet, og siktemålet for prosjektet var å granske sosiale forskjeller i Oslo-målet. TAUS består av uformelle intervjuer med Oslo-folk i alderen 15-17 og 34-75 år. Materialet utgjør ca. 24 timer opptak, og er på ca. 246 000 ord."

På basis av disse korpusene har jeg laget de følgende matematiske modellene for dativalternering, og sjekket deres prediksjoner, ved følgende generelle R-kommandoer:

```
txtdata <- read.table("korpus.txt", header=T)

numdata <- transform(txtdata, logdiff = log(lrec/lth)))
```

```
modellnavn <- glm(real ~ class + accrec + accth + prorec +
     proth + defrec + defth + animrec + animth + logdiff,
     family=binomial, data=numdata)

summary(modellnavn)

table(numdata$real, predict(modellnavn) > 0.0)
```

Disse kommandoene leser først inn de kodete korpusdataene, tar logaritmen av de relative fraselengdene, bygger selve modellen ved å beregne parameterkoeffisientene, gir en summarisk oversikt over de beregnede verdiene og deres signifikans, og krysstabulerer til slutt de virkelige realiseringene av recipient (NP eller PP) mot modellens prediksjoner for det samme. Grunnlaget for prediksjonene er her all modellinput tatt i sin totalitet, gitt som en n×p-matrise X med inputverdier, der n er antall observasjoner (korpustreff) og p er antall modellparametre (faktorer), og en p×1-matrise β med beregnede parameterkoeffisienter. Modellformelen for en logistisk regresjonsmodell for sannsynligheten for realisering av recipient som PP blir da: $f(z) = 1/(1+e^{-z})$ der $z = X_i\beta$. Modellen beregner altså sannsynligheten for at recipient realiseres som enten første NP i en dobbelobjekt-konstruksjon V NP NP (når $f(z)<0,5$ og denotert som 0) eller som en PP i en konstruksjon av typen V NP PP (når $f(z)>0,5$ og denotert som 1). Skillet er altså skarpt, slik at enhver sannsynlighet større enn 0,5 vil gi opphav til en PP-prediksjon.

I de resulterende modellene vil typisk ikke alle parametrene være statistisk signifikante, og det vil følgelig være ønskelig å remodellere med disse parametrene utelatt. Dette vil også endre koeffisientverdiene for de gjenværende parametrene, men skulle ikke resultere i nye insignifikante parametre. Det vil også være samvariasjon (korrelasjon) mellom noen parametre, og R inneholder prosedyrer for å beregne og vise disse korrelasjonene. Videre er det en viss fare for såkalt 'overfit', at modellene bare memorerer input og ikke representerer noen genuin generalisering som kan brukes til å predikere ukjente data. For å undersøke om dette er tilfelle, kan man teste modellene mot andre, beslektede datasett. For eksempel kan en modell som bygger på et skriftspråkkorpus testes på et talespråkkorpus eller motsatt. En annen mulig metode er å remodellere gjentatte ganger mot tilfeldige deler av et korpus (for eksempel 100 ganger mot 85 % av korpuset valgt tilfeldig hver gang), og bruke gjennomsnittet som ny modell. Dette garanterer at et lite antall ekstremverdier i datasettet ikke resulterer i en "skjev" modell.

De ferdige modellene og deres prediksjoner danner grunnlaget for den videre drøftingen av hvilke faktorer som faktisk påvirker dativalterneringen i norsk, hvordan de samvirker, og hvilke likheter og forskjeller som finnes i forhold til tilsvarende undersøkelser i andre språk, nærmere bestemt ulike varianter av engelsk.

# 4 Resultater

I det følgende presenteres resultatene av ulike modelleringer:

a) Bruk av *gi* i norske aviser.

b) Bruk av *gi* i norsk tale.

c) Bruk av noen ditransitive verb i Aftenposten.

Hver modell gjenomgås i egne delkapitler.

Kapittelet begynner med en oversikt over de rådataene som ligger til grunn for modelleringene, og fortsetter med en detaljert gjennomgang av hver enkelt modell. Jeg har tatt med komplette utskrifter fra modelleringsverktøyet R, slik at det skal være mulig å etterprøve holdbarheten av modellene, og selv kunne foreta alternative beregninger av modellsignifikans, etc.

Modellene som gjennomgås går fra separate modeller for skrift og tale, til modeller som kombinerer begge typer inndata. Alle modellene bruker de samme antatt uavhengige parametrene. Den nest siste modellen i rekken ser i tillegg på interaksjonseffekter mellom noen av parametrene, mens den siste introduserer en såkalt mixed effect modell.

## 4.1 Oversikt over rådata

Datamaterialet som ligger til grunn for modelleringene for verbet *gi* i norske aviser og tale, kan oppsummeres i følgende tabeller:

### 4.1.1 Kilder versus realisering

| gi | kilde | NP | PP | PP % |
|---|---|---|---|---|
| tale | Big Brother | 90 | 11 | 10,9 % |
| | Nota | 59 | 20 | 25,3 % |
| | Taus | 17 | 3 | 15,0 % |
| | | | | |
| skrift | Adresseavisen | 180 | 39 | 17,8 % |
| | Aftenposten | 289 | 56 | 16,2 % |
| | Bondebladet | 20 | 9 | 31,0 % |
| | Bergens Tidende | 1097 | 251 | 18,6 % |
| | Dagbladet | 1 | 1 | 50,0 % |
| | Det Nye | 83 | 6 | 6,7 % |
| | Familien | 107 | 7 | 6,1 % |
| | Hennes | 53 | 4 | 7,0 % |
| | HP | 1 | 0 | 0,0 % |
| | KK | 23 | 0 | 0,0 % |
| | Motor | 19 | 3 | 13,6 % |
| | Stavanger Aftenblad | 137 | 31 | 18,5 % |
| | Universitas | 1 | 0 | 0,0 % |
| | VG | 133 | 17 | 11,3 % |
| | Vårt Land | 133 | 26 | 16,4 % |

Tabell: Delkorpus

Vi ser her at delkorpusene har svært ulik andel PP-er både for tale og skrift. For skrift kan en merke seg at de typiske abonnementsavisene (Adresseavisen, Aftenposten, Bergens Tidende, Stavanger Aftenblad og Vårt Land) bruker PP-er mer enn dobbelt så ofte som de typiske ukebladene (Det Nye, Familien, Hennes), med henholdsvis 16-19 % mot 6-7 %, mens VG ligger midt imellom disse to gruppene, med ca 11 %. Dette kan tyde på at stil, eller grad av formalitet, er ytterligere en faktor som kunne inkluderes i modelleringen.

### 4.1.2 Faktorer versus realisering

| **gi** | | **skrift** | | **tale** | |
|---|---|---|---|---|---|
| faktor | verdi | NP | PP | NP | PP |
| Aksesserbarhet: recipient | acc | **49,9 %** | 3,0 % | **80,0 %** | 5,5 % |
| | notacc | 33,6 % | 13,5 % | 3,0 % | 11,5 % |
| | | | | | |
| Aksesserbarhet: theme | acc | 1,3 % | *0,9 %* | 9,5 % | *11,5 %* |
| | notacc | ***82,2 %*** | 15,6 % | ***73,5 %*** | 11,0 % |
| | | | | | |
| Definitthet: recipient | def | **78,1 %** | 11,6 % | **82,0 %** | 14,0 % |
| | notdef | 5,4 % | 5,0 % | 1,0 % | 3,0 % |
| | | | | | |
| Definitthet: theme | def | 15,9 % | *3,6 %* | 20,5 % | *11,0 %* |
| | notdef | ***67,6 %*** | 12,9 % | ***62,5 %*** | 6,0 % |
| | | | | | |
| Pronominalitet: recipient | pro | **36,6 %** | 0,7 % | **75,0 %** | 5,5 % |
| | notpro | 46,9 % | 15,8 % | 8,0 % | 11,5 % |
| | | | | | |
| Pronominalitet: theme | pro | 0,5 % | *0,5 %* | 2,5 % | *9,5 %* |
| | notpro | ***83,0 %*** | 16,0 % | ***80,5 %*** | 7,5 % |
| | | | | | |
| Animathet: recipient | anim | **58,0 %** | 7,5 % | **79,0 %** | 14,5 % |
| | notanim | 25,4 % | 9,0 % | 4,0 % | 2,5 % |
| | | | | | |
| Animathet: theme | anim | 0,4 % | *0,1 %* | 0,0 % | *0,5 %* |
| | notanim | ***83,1 %*** | 16,4 % | ***83,0 %*** | 16,5 % |
| | | | | | |
| Relativ lengde | rec>theme | 8,1 % | 8,8 % | 0,0 % | 5,0 % |
| | rec<=theme | **75,4 %** | 7,7 % | **83,0 %** | 12,0 % |
| | | | | | |
| Semantisk klasse | abstract | 72,2 % | 11,7 % | 50,5 % | 6,0 % |
| | comm | 3,7 % | 1,7 % | 6,5 % | 0,5 % |
| | transfer | 7,6 % | 3,1 % | 26,0 % | 10,5 % |

Tabell: Rådata

Intuisjonen og forventningen som understøttes av tallene ovenfor er at aksesserbare, definitte, pronominale og/eller animate recipienter, som dessuten er kortere enn theme, vil bli realisert som NP, ikke PP. De **uthevede** tallene i tabellen ovenfor synes i stor grad å bekrefte dette, med et mulig unntak for pronominalitet av recipienten. Videre er det forventet og intuitivt

rimelig at aksesserbar, definitt, pronominal og/eller animat theme vil bli realisert rett etter verbet, slik at recipienten da blir realisert som PP. De *kursiverte* tallene i tabellen ovenfor kan imidlertid ikke bekrefte denne forventningen, rett og slett fordi andelene av theme med disse egenskapene er for små i datagrunnlaget. Indirekte blir forventningen likevel bekreftet ved å se på theme med motsatte egenskaper, ***kursiverte uthevede tall***, som i følge denne logikken burde vise en klar tendens til at recipienten realiseres som NP, noe de også gjør. Tallene understøtter altså den teoretiske forventningen om at recipient eller theme med høy prominens på de undersøkte hierarkiene, blir realisert i en prominent syntaktisk posisjon i setningen. Tallene kan imidlertid ikke uten videre si noe om hvordan de undersøkte faktorene samvirker i å produsere en gitt realisering. Dette kan bare en konkret matematisk modell gjøre, som i de følgende delkapitlene.

Når det gjelder forskjeller og likheter mellom skrift og tale, er det her verdt å merke seg at en mindre andel av recipientene er aksesserbare (aksesserbarhet: recipient = acc) i skrift enn i tale, med henholdsvis 49,9 % mot 80 %. Videre er en mye mindre andel av recipientene pronominale (definitthet: recipient = pro) i skrift enn i tale, med henholdsvis 36,6 % mot 75 % Dessuten er en noe mindre andel av recipientene animate (animathet: recipient = anim) i skrift enn i tale, med henholdsvis 58 % mot 79 %. Til sist er det vanligere å bruke verbet *gi* i betydningen fysisk overføring (semantisk klasse = transfer) i tale enn i skrift, med henholdsvis 26 % mot 7,6 %. Disse funnene er i tråd med hva en kan forvente, for så vidt som recipienten i tale oftere er et personlig pronomen, og således både aksesserbar, pronominal og animat.

Når recipienten er aksesserbar er det en tydelig preferanse for realisering som NP både i skrift og tale. I skrift skjer dette minst 16 ganger (49,9 / 3,0) så ofte som den realiseres som PP, i tale minst 14 ganger (80,0 / 5,5) så ofte. Ikke-aksesserbare recipienter viser et mye mer forvirrende mønster. I skrift NP-realiseres recipienten minst dobbelt (33,6 / 13,5) så ofte som den realiseres som PP, mens i tale vil motsatt recipienten PP-realiseres nesten fire ganger (11,5 / 3,0) så ofte som den NP-realiseres. I skrift er totalt sett om lag halvparten av alle recipienter realisert som NP og aksesserbare, mens om lag en tredjedel er ikke-aksesserbare og likevel realisert som NP. I tale er derimot fire av fem recipienter både realisert som NP og aksesserbare. Ikke-aksesserbar recipient forekommer i det hele tatt mye oftere i skrift enn i tale. Det bør understrekes at aksesserbarhet er vanskelig å bestemme sikkert ut fra den begrensede konteksten som korpus-treffene gir, slik at denne faktoren er den mest usikre hva

innkodingen av materialet angår. I tale motvirkes dette av at recipienten der mye oftere er et personlig pronomen, og altså lettere å bestemme hva angår aksesserbarhet. Den utydelige korrelasjonen mellom ikke-aksesserbarhet og realisering som NP eller PP kan dermed være en konsekvens av problemer ved kodingen.

Aksesserbarhet eller ikke av theme kunne ha hatt samme problem med kodingen som aksesserbarhet av recipient, men viser likevel både i tale og skrift et klart mønster. I fire av fem (for skrift) eller tre av fire (for tale) tilfeller vil theme være ikke-aksesserbar og recipient samtidig realisert som en NP. Når theme er aksesserbar er det ingen bestemt preferanse for NP eller PP verken i skrift eller tale, mens når theme er ikke-aksesserbar vil recipienten i skrift NP-realiseres minst fem ganger så ofte som den PP-realiseres, og i tale mer enn seks ganger så ofte. Aksesserbar theme forekommer dessuten totalt sett sjeldnere i skrift enn i tale.

Definitthet av recipient viser et like klart mønster. I fire av fem tilfeller, så vel i skrift som i tale, vil en recipient være både definitt og realisert som NP. Videre viser tallene at en definitt recipient vil NP-realiseres om lag seks ganger så ofte som den PP-realiseres, både i skrift og i tale. Indefinitte recipienter viser ingen bestemt preferanse for realisering som NP eller PP i skrift, men tenderer til å PP-realiseres i tale.

Definitthet av theme viser også et klart mønster, om enn mindre kategorisk. I to av tre tilfeller, både i skrift og i tale, vil recipienten realiseres som NP og theme være indefinitt. Indefinitt theme gir i skrift NP-realisering av recipienten minst fem ganger så ofte som PP-realisering, og minst ti ganger så ofte i tale. Definitt theme viser imidlertid også preferanse for å realisere recipienten som NP, med henholdsvis minst fire ganger så ofte som PP-realisering i skrift, og nesten dobbelt så ofte i tale.

Pronominalitet av recipient viser i likhet med aksesserbarhet av recipient et noe uklart totalbilde. I om lag en tredjedel av tilfellene vil recipient være pronominal og realisert som NP, mens den er ikke-pronominal og likevel realisert som NP i om lag halvparten av tilfellene. Hvis den er pronominal vil den imidlertid nesten aldri realiseres som PP i skrift, faktisk vil den da minst 40 ganger så ofte NP-realiseres. Også i tale vil en pronominal recipient realiseres som NP nesten 15 ganger så ofte som den realiseres som PP. I tale vil tre av fire recipienter være pronominale og realisert som NP, og det forekommer også en viss realisering som PP selv om recipienten er pronominal. Ikke-pronominale recipienter viser

imidlertid også en klar preferanse for NP-realisering i skrift, med om lag tre ganger så ofte som PP-realisering, mens de ikke viser noen klar preferanse for verken NP eller PP i tale.

Pronominalitet av theme viser klart at i om lag fire av fem tilfeller er theme ikke-pronominal og recipient realisert som NP, og theme er nesten aldri pronominal i skrift. Theme som er pronominal viser i tale en preferanse for PP-realisering av recipienten knapt fire ganger så ofte som NP-realisering, mens ikke-pronominal theme preferer NP-realisering av recipienten mer enn fem ganger så ofte som PP-realisering i skrift, og minst ti ganger så ofte i tale.

Animathet av recipient viser et tydeligere totalmønster enn aksesserbarhet og pronominalitet. I skrift vil seks av ti recipienter være animate og realisert som NP, mens en av fire vil være ikke-animate og likevel realisert som NP. I tale vil fire av fem recipienter være animate og realisert som NP. Dette betyr at animate recipienter i skrift NP-realiseres nesten åtte ganger så ofte som de realiseres som PP, og i tale minst fem ganger så ofte. Imidlertid viser også ikke-animate recipienter en viss preferanse for NP-realisering i skrift, med nesten tre ganger så ofte som PP-realisering, mens det i tale ikke synes å være noen preferanse noen vei.

Animathet av theme viser tydelig at i vel fire av fem tilfeller er theme ikke-animat og recipient realisert som NP, og at theme nesten aldri er animat verken i skrift eller i tale. Dermed kan man ikke fastslå noen preferanser noen vei for animat theme, mens ikke-animat theme foretrekker NP-realisering av recipienten om lag fem ganger så ofte som PP-realisering.

Relativ lengde viser i tre av fire tilfeller (i skrift) eller fire av fem tilfeller (i tale) at recipienten er realisert som NP og samtidig theme er lengre enn (eller like lang) som den. I tale forekommer det dessuten nesten aldri at recipienten er lengre enn theme og likevel realisert som NP. I skrift viser det seg da ingen preferanse for verken NP eller PP, mens når recipienten er kortere eller lik theme vil recipienten i skrift nesten ti ganger så ofte NP-realiseres som den vil PP-realiseres, og i tale om lag sju ganger så ofte.

Hvis man slår sammen dataene for både skrift og tale i en fil, gir en krysstabulering av treff (av totalt 2927) med oppgitt recipientlengde (vertikalt) og themelengde (horisontalt) følgende matrise for antall forekomster:

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 722 | 563 | 384 | 170 | 118 | 53 | 39 | 24 | 21 | 3 | 3 | 3 | 5 | 1 | 4 |
| **2** | 119 | 80 | 74 | 24 | 24 | 10 | 5 | 2 | 2 | 1 | 1 | 0 | 0 | 0 | 0 |
| **3** | 71 | 50 | 56 | 7 | 9 | 4 | 3 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| **4** | 27 | 21 | 11 | 12 | 5 | 4 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **5** | 21 | 20 | 13 | 5 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **6** | 13 | 15 | 4 | 2 | 2 | 0 | 1 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| **7** | 8 | 11 | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **8** | 4 | 9 | 6 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **9** | 4 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **10** | 5 | 4 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **11** | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **12** | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **13** | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **14** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Tabell: Fraselengder

Dette viser klart at recipienten som oftest er kortere enn theme i dette materialet.

Semantisk klasse viser at recipienten oftest realiseres som NP i alle tre bruksmåter. I abstrakt bruk skjer dette i skrift om lag seks ganger så ofte som PP-realisering, og i tale om lag åtte ganger så ofte. For kommunikativ bruk er det mest påfallende at det i tale sjelden kommer til realisering som PP i det hele tatt, mens det for fysisk overføring er mest påfallende at denne bruksmåten opptrer mye hyppigere i tale enn i skrift, som nevnt tidligere, og da uansett realisering av recipienten som NP eller PP.

En del av likhetene mellom disse mange tallene skyldes underliggende korrelasjoner. Det er selvfølgelig en viss samvariasjon mellom for eksempel definitthet, pronominalitet og animathet for både recipient og theme. Det er likevel klart bare ut fra disse tallene at hver enkelt faktor potensielt kan bidra til å avgjøre valg av NP eller PP for recipienten, og dermed måtte inngå i en faktoranalyse. Dette gjøres i de følgende delkapitlene.

## 4.2 Bruk av verbet gi i norske aviser

Datamaterialet her omfatter totalt 2727 treff for ditransitiv bruk av verbet. Modellen gav følgende oppsummering og parameterkoeffisienter (resultater i `Courier New`, mine kommentarer i Times New Roman):

```
> summary(modelS)
Call:
glm(formula = real ~ class + accrec + accth + prorec + proth +
    defrec + defth + animrec + animth + logdiff, family =
    binomial, data = numavis)
```

Dette gjentar bare kommandoen som produserte modellen. Modellformelen angir at den avhengige variabelen realisering (real) er en funksjon av de uavhengige variablene klasse (class), aksesserbarhet, pronominalitet, definitthet og animathet av henholdsvis recipient (accrec, prorec, defrec og animrec) og theme (accth, proth, defth og animth), samt relativ fraselengde av recipient og theme (logdiff). Modellen er spesifisert som binomial, og underforstått som logistisk. Kodete inndata hentes fra filen 'numavis.txt'.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3418  -0.5257  -0.2178  -0.1049   3.2267
```

Disse tallene oppsummerer at hovedtyngden av treffene (mer enn 3Q, altså mer enn 75 %) er realisert som NP (negative verdier).

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.254784   0.161730  -1.575   0.1152
classc       1.095345   0.231778   4.726 2.29e-06 ***
classt       1.370081   0.188350   7.274 3.49e-13 ***
```

```
accrec      -0.430209   0.171675  -2.506   0.0122 *
accth        0.664257   0.398641   1.666   0.0957 .
prorec      -1.628399   0.277287  -5.873 4.29e-09 ***
proth        2.310037   0.530201   4.357 1.32e-05 ***
defrec      -0.846946   0.163805  -5.170 2.34e-07 ***
defth        0.007045   0.166906   0.042   0.9663
animrec     -0.759672   0.134359  -5.654 1.57e-08 ***
animth       1.185844   0.998830   1.187   0.2351
logdiff      1.165373   0.081763  14.253  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

Dette gir estimerte parameterkoeffisienter. Legg i kolonnen merket ’Estimate’ merke til at hva recipienten angår er aksesserbarhet (accrec), pronominalitet (prorec), definitthet (defrec) og animathet (animrec) alle sammen negative, og trekker i samme retning: at aksesserbar, pronominal, definitt og/eller animat recipient realiseres som NP (negative verdier). Hva theme angår er aksesserbarhet (accth), pronominalitet (proth), definitthet (defth), og animathet (animth) alle sammen positive, og trekker også i samme retning: at aksesserbar, pronominal, definitt og/eller animat theme medvirker til at recipienten realiseres som PP (positive verdier). Med unntak av aksesserbarhet, definitthet og animathet av theme er alle disse parametrene klart signifikante: stjernekodingen i høyre kolonne viser at de har minst en stjerne og at sannsynligheten for at de egentlig er lik 0 (og altså ikke bidrar i noen retning) dermed er mindre enn 0,05. (De er signifikante med $p < 0,05$.) Intercept-verdien alene viser korrekt nok at en statisk modell uten noen parametre forutsier realisering som NP (negativ verdi), mens relativ fraselengde (logdiff) ikke overraskende tyder på at recipient som er lengre enn theme realiseres som PP (positiv verdi).

De signifikante toverdige koeffisientverdiene har følgende tolkning: Aksesserbare recipienter har i seg selv mer enn 50 % så høy sjanse til å realiseres som NP som ikke-aksesserbare ($e^{0,430209} = 1,5$). Pronominale recipienter har over fem ganger så høy sjanse til å realiseres som NP som ikke-pronominale ($e^{1,628399} = 5,1$). Definitte recipienter har over dobbelt så høy sjanse

til å realiseres som NP som indefinitte ($e^{0,846946}$ = 2,3). Animate recipienter har over dobbelt så høy sjanse til å realiseres som NP som ikke-animate ($e^{0,759672}$ = 2,1). Pronominalt theme har mer enn ti ganger så høy sjanse til at recipienten realiseres som PP som ikke-pronominalt theme ($e^{2,310037}$ = 10,1).

```
    Null deviance: 2442.8  on 2726  degrees of freedom
Residual deviance: 1646.5  on 2715  degrees of freedom
AIC: 1670.5

Number of Fisher Scoring iterations: 6
```

Dette er tekniske verdier som i hovedsak angir at en statisk modell uten parametre (null-modellen) er dårligere tilpasset datamaterialet enn den fulle modellen. Jo større forskjell det er mellom 'null deviance' og 'residual deviance', jo bedre kan modellparametrene forklare variansen i det underliggende datamaterialet. Her er det tydelig at en betydelig andel av variansen ikke kan forklares av modellen, siden 'deviance' bare har blitt redusert med om lag en tredjedel.

Det overraskende resultatet i denne modellen, er at definitthet av theme (defth) ikke synes å spille noen signifikant rolle. Dette er ikke tilfelle for noen av Bresnans mange ulike modeller for ulike varianter av engelsk. Derimot utelater også Bresnan animathet av theme (animth) fra de endelige modellene, siden denne parameteren ikke er signifikant.

Når det gjelder hva modellen predikerer, er resultatene mindre tilfredsstillende. Følgende er en krysstabulering av faktisk realisering (vertikalt) versus prediksjon (horisontalt: TRUE hvis PP predikeres, FALSE ellers):

```
> table(numavis$real,predict(modelS)>0.0)
```

| | FALSE | TRUE | % korrekt predikert |
|---|---|---|---|
| NP | 2191 | 86 | 96,2 % |
| PP | 268 | 182 | 40,4 % |

Modellen har god prediksjonskraft for NP, men tar systematisk feil for PP. Totalt gir dette at 87,0 % (2191 + 182 = 2373 av totalt 2727) predikeres korrekt. Tatt i betraktning av at det totalt er 83,5 % (2191 + 86 = 2277 av totalt 2727) NP-er i datagrunnlaget, gir dette at modellen samlet sett bare er marginalt bedre enn en "modell" som alltid velger NP, uansett.

Det er altså bare 16,5 % (268 + 182 = 450 av totalt 2727) PP-er i datagrunnlaget, og det virker som om noen av de parameter-kombinasjonene som i modellen trekker i retning NP, i virkeligheten like gjerne kan gi PP. Spørsmålet er om dette er et resultat av at modellen bygger på et skriftspråkkorpus hvor man kan tenke seg at den ekstra bearbeidingen språkbrukerne legger i å formulere seg skriftlig overstyrer deres valg av dativalternativ i naturlig tale, om det gjenspeiler særegne forhold ved verbet *gi*, eller om det kan finnes andre forklaringer. For å undersøke dette har jeg sett på noen andre datasett og modeller.

## 4.3 Bruk av verbet gi i norsk tale

Datamaterialet omfatter her totalt bare 200 treff for ditransitiv bruk av verbet. Modellen gav følgende oppsummering og parameterkoeffisienter (resultater i `Courier New`, mine kommentarer i Times New Roman):

```
> summary(modelT)

Call:
glm(formula = real ~ class + accrec + accth + prorec + proth +
    defrec + defth + animrec + animth + logdiff, family =
    binomial, data = numtale)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5265  -0.1564  -0.0646  -0.0170   3.6393

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    4.5111     2.9350   1.537  0.12429
classc         0.2585     1.9757   0.131  0.89590
classt         1.1141     0.8394   1.327  0.18444
accrec        -1.3678     1.2741  -1.074  0.28300
accth          1.5064     1.1134   1.353  0.17605
```

```
prorec        -1.2650     1.2252  -1.032  0.30187
proth          3.0673     1.4659   2.092  0.03640 *
defrec        -2.1631     1.6523  -1.309  0.19048
defth         -1.6767     1.2299  -1.363  0.17280
animrec       -3.2095     2.5648  -1.251  0.21081
animth        13.4164  2399.5450   0.006  0.99554
logdiff        3.8600     1.4391   2.682  0.00731 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 182.354  on 199  degrees of freedom
Residual deviance:  55.074  on 188  degrees of freedom
AIC: 79.074

Number of Fisher Scoring iterations: 15
```

Modellen bygger på et for lite datamateriale til at parameterkoeffisientene kan estimeres med tilstrekkelig grad av signifikans, noe som vises av at bare pronominalitet av theme (proth) og relativ fraselengde (logdiff) er signifikante med feilsannsynlighet *mindre* enn 0,05. Men siden den store forskjellen mellom ’null deviance’ og ’residual deviance’ tydelig viser at modellen i stor grad kan forklare variansen i det underliggende datamaterialet, kan vi anta at den også viser tendensene i dette materialet. Vi ser at aksesserbarhet, pronominalitet, definitthet og animathet av recipient alle er negative og trekker i samme retning som i forrige modell (recipient realiseres som NP), mens aksesserbarhet, pronominalitet og animathet av theme er positive, og, som i forrige modell, gir at recipienten vil tendere mot å realiseres som PP. Forskjellen fra forrige modell er definitthet av theme (defth), som der var positiv. Siden defth uansett ikke er signifikant i noen av modellene, kan man se bort fra dette.

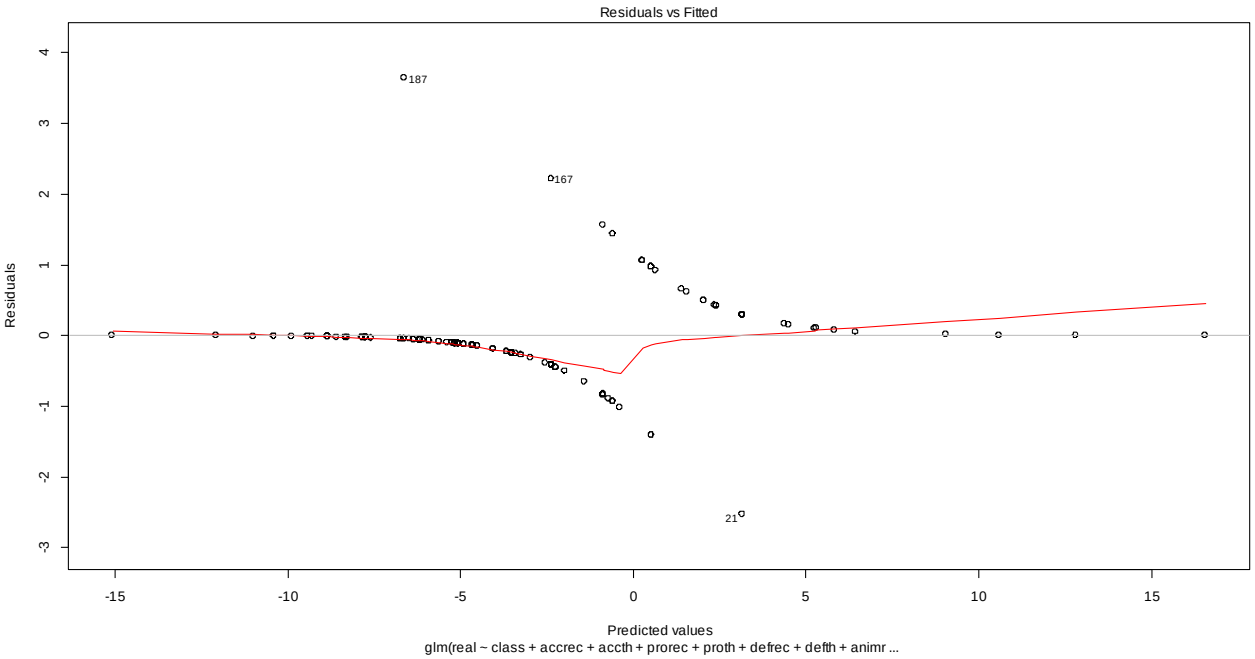
Modellen gir følgende prediksjoner, krysstabulert mot faktisk realisering:

```
> table(numtale$real,predict(modelT)>0.0)
```

| | FALSE | TRUE | % korrekt predikert |
|---|---|---|---|
| NP | 163 | 3 | 98,2 % |
| PP | 4 | 30 | 88,2 % |

Vi ser her at modellen faktisk predikerer resultatene bedre enn forrige modell. 98,2 % av NP-ene predikeres korrekt, men dessuten også 88,2 % av PP-ene. Totalt gir dette at modellen predikerer 96,5 % av resultatene korrekt (163 + 30 = 193 av 200).

Følgende graf viser hvordan treffene fordeler seg i forhold til modellens prediksjoner.




Graf: Gi (tale)

Hvert punkt i grafen representerer ett treff. Dets horisontale plassering angir predikert realisering (NP for negative horisontale akseverdier, PP for positive), mens dets vertikale avstand til den prikkede rette linjen angir treffets faktiske avvik (’deviance residual’) fra denne predikerte verdien. Den heltrukne linjen er den kurven som i henhold til modelleringsalgoritmen passer best til punktene.

Noen få treff predikeres feil i forhold til kodet realisering, for eksempel nummer 187 øverst til venstre:

”faren min er litt sånn derre Ø altfor snill Ø så han hadde gitt Ø alt Ø vi hadde Ø til dem”

Dette treffet er kodet som ”PP gi t gi.t 1 0 1 0 1 0 1 0 1 3 Nota”: animat, definitt, pronominal, og aksesserbar recipient (*dem*); ikke-animat, ikke-definitt, ikke-pronominal og ikke-

aksesserbar theme (*alt vi hadde*: merk at *alt* ikke regnes som pronominal og definitt her); *gi* brukt om fysisk overføring. Modellen predikerer treffet som NP til tross for at det er en PP.

Nummer 21 nederst til høyre er litt annerledes:

> NP gi t gi.t 1 0 1 1 0 1 0 1 1 1 BB
>
> Monica sier:
>
> ”tilby å Ø bare sitter der og ser på (uforståelig) gi n til e Rodney nå ja men i fj- går”

Dette treffet predikeres som PP til tross for at det er *kodet* som en NP. Det må regnes som en suksess for modellen at dette faktisk representerer en *feilkoding*, siden treffet vitterlig bruker en PP (*til Rodney*). Recipienten (*Rodney*) er kodet som animat, definitt, ikke-pronominal og ikke-aksesserbar (det siste kan diskuteres i sin kontekst). Theme (*n*) er kodet som ikke-animat, definitt, pronominal og aksesserbar. Verbet *gi* er brukt om fysisk overføring.

Totalt er det i dette talebaserte datamaterialet 17,0 % PP-er, altså praktisk talt det samme som i det foregående avis-materialet. Det ser altså ut til at verbet *gi* brukes likt i tale og skrift hva realisering av recipient angår, og at forklaringen på hvorfor skriftsspråks-modellen i liten grad greide å forutsi realisering som PP, må søkes annetsteds. Kanskje har det rett og slett med stil, eller grad av formalitet, å gjøre. For å undersøke om det samme er tilfelle med andre verb, så jeg også på det Aftenposten-materialet jeg hadde samlet (og kodet) før jeg gikk over til bare å se på verbet *gi*.

## 4.4 Bruk av noen ditransitive verb i Aftenposten

Datamaterialet omfatter her totalt 511 treff for ditransitiv bruk av verbene:

- bringe, by, gi, leie, låne, overlate, overrekke, rekke, selge, sende, servere, skaffe, skjenke, tilby, tildele, tilegne, betale, bevilge, betro, fortelle, love og meddele.

Modellen gav følgende oppsummering og parameterkoeffisienter (resultater i `Courier New`, mine kommentarer i Times New Roman):

```
> summary(modelA)
Call:
glm(formula = real ~ class + accrec + accth + prorec + proth +
    defrec + defth + animrec + animth + logdiff, family =
    binomial, data = numaften)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9372  -0.6761  -0.2957   0.5873   2.6114

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.2167     0.3443  -0.629 0.529059
classc        1.5278     0.4076   3.749 0.000178 ***
classt        1.0334     0.2716   3.805 0.000142 ***
accrec       -0.9724     0.2913  -3.338 0.000843 ***
accth         0.7105     0.3815   1.862 0.062549 .
prorec       -1.3515     0.4737  -2.853 0.004326 **
proth         0.7687     1.2099   0.635 0.525210
defrec       -0.5676     0.3271  -1.735 0.082712 .
defth         0.2561     0.3160   0.811 0.417641
animrec      -0.5752     0.2542  -2.263 0.023631 *
animth        0.6867     1.5447   0.445 0.656658
logdiff       0.6610     0.1392   4.748 2.05e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 603.95  on 510  degrees of freedom
Residual deviance: 425.67  on 499  degrees of freedom
AIC: 449.67

Number of Fisher Scoring iterations: 6
```

Også denne modellen bygger på et for lite datamateriale, men i alle fall er nå aksesserbarhet (accrec), pronominalitet (prorec), animathet (animrec) av recipient, og relativ fraselengde (logdiff), alle bedre enn sannsynlighet mindre enn 0,05 for *ikke* å være signifikante. Videre ser vi at alle parametrene trekker i samme retning som i den første skriftspråksmodellen, den

for *gi* alene. Dette er for så vidt ikke overraskende siden *gi* står for brorparten av treffene i dette materialet også (ca 350 av 511 treff). Derimot tyder den relativt beskjedne reduksjonen i 'deviance' på at en stor del av variansen i datagrunnlaget ikke kan forklares av modellen.

Modellens prediksjoner er som følger, krysstabulert mot faktisk realisering:

```
> table(numaften$real,predict(modelA)>0.0)
```

| | FALSE | TRUE | % korrekt predikert |
|---|---|---|---|
| NP | 334 | 35 | 90,5 % |
| PP | 67 | 75 | 52,8 % |

90,5 % av NP-ene predikeres korrekt, men bare 52,8 % av PP-ene. Totalt gir dette 80,0 % korrekt, i et materiale som består av 27,8 % PP-er og 72,2 % NP-er. Andelen PP-er i dette faktiske materialet er altså betydelig høyere enn for *gi* alene, men den totale prediksjons-kraften til modellen er også her bare litt bedre enn for en "modell" som alltid velger NP. Det er ut fra dette materialet tydelig at andre verb realiserer recipient som PP mye oftere enn *gi*, men antall treff for disse andre verbene er for lite til at det er mulig å si noe sikkert om hvordan de hver for seg oppfører seg.

Denne modellen egner seg for sammenlikning med Bresnan et als (2007, s. 81) første modell for dativalternering i engelsk, basert på SWITCHBOARD-korpuset av telefonsamtaler. Bresnan et al gir følgende formel for valg av alternativ (modell A):

Probability(Response = 1) = $1/(1+e^{-X\beta})$, der
Xβ = 0,95 - 1,34{c} + 0,53{f} - 3,90{p} + 0,96{t}
+0,99{accessibility of recipient = nongiven}
-1,1{accessibility of theme = nongiven}
+1,2{pronominality of recipient = nonpronoun}
-1,2{pronominality of theme = nonpronoun}
+0,85{definiteness of recipient = indefinite}
-1,4{definiteness of theme = indefinite}
+2,5{animacy of recipient = inanimate}
+0,48{person of recipient = nonlocal}
-0,03{number of recipient = plural}

+0,5{number of theme = plural}

-0,46{concreteness of theme = nonconcrete}

-1,1{parallelism = 1}

-1,2*length difference (log scale)

Her er {…} = 1 hvis betingelsen holder, og {…} = 0 hvis den ikke holder.

Tilsvarende gir min modell følgende formel:

Probability(Response = 1) = $1/(1+e^{-X\beta})$, der

$X\beta$ = -0,22 + 1,53{c} + 1,03{t}

-0,97{accessibility of recipient = given}

+0,71{accessibility of theme = given}

-1,35{pronominality of recipient = pronoun}

+0,77{pronominality of theme = pronoun}

-0,57{definiteness of recipient = definite}

+0,26{definiteness of theme = definite}

-0,57{animacy of recipient = animate}

+0,69{animacy of theme = animate}

+0,66*relative length (log scale)

Forskjellen i fortegn for de binære faktorene i disse to formlene skyldes at Bresnan et al har kodet fravær av faktoren (altså nongiven, nonpronoun, indefinite og inanimate) som TRUE (1), mens jeg motsatt har kodet tilstedeværelse av faktoren (altså given, pronoun, definite og animate) som TRUE (1), i tråd med hva jeg er vant til fra matematisk logikk og programmering. Forskjellen i fortegn for length difference (Bresnan et al) kontra relative length (min modell) skyldes i prinsippet at Bresnan et al har kodet for lengde av theme minus lengde av recipient, mens jeg motsatt har kodet for lengde av recipient minus lengde av theme, men kompliseres også av at det er brukt to ulike måter å log-transformere uttrykket. Forskjellen i fortegn på intercept og (til dels) semantisk klasse må være en konsekvens av alle disse andre ulikhetene.

Hva valg av faktorer angår, har jeg i motsetning til Bresnan et al ikke kodet for person og tall for recipient. Videre har jeg kodet for animathet av theme, i motsetning til Bresnan et al som i stedet har kodet for tall og konkrethet av theme. Det er verdt å merke seg at disse fire ekstra

faktorene (person og tall for recipient, tall og konkrethet av theme) ikke bidrar i noen stor grad i Bresnan et als modellformel. Bresnan et al har derimot ytterligere en viktig faktor, parallellisme, som jeg ikke fant å kunne ta med, da den krever for mye kontekst for å kunne bestemme. Parallellisme er en priming-faktor, for så vidt som nylig brukte strukturer tenderer til å bli gjentatt.

Den relative vektingen av de faktorene som er felles for Bresnan et als og min modellformel, viser at engelsk og norsk oppfører seg ganske likt, men det er noen åpenbare forskjeller. Det synes som theme generelt har mindre betydning i norsk enn engelsk, og engelsk ser ut til å legge betydelig større vekt på animathet av recipienten enn norsk. At norsk legger mindre vekt på relativ lengde, kan skyldes morfologiske ulikheter, og derav følgende forskjeller i telling av ord. (For eksempel at det er færre sammenskrevne ord i engelsk og dermed kanskje flere ortografiske ord i en frase enn i den tilsvarende norske frasen, og at norsk har etterstilt bestemt artikkel i motsetning til engelsk *the*, slik at en norsk frase også her blir kortere enn den tilsvarende engelske.)

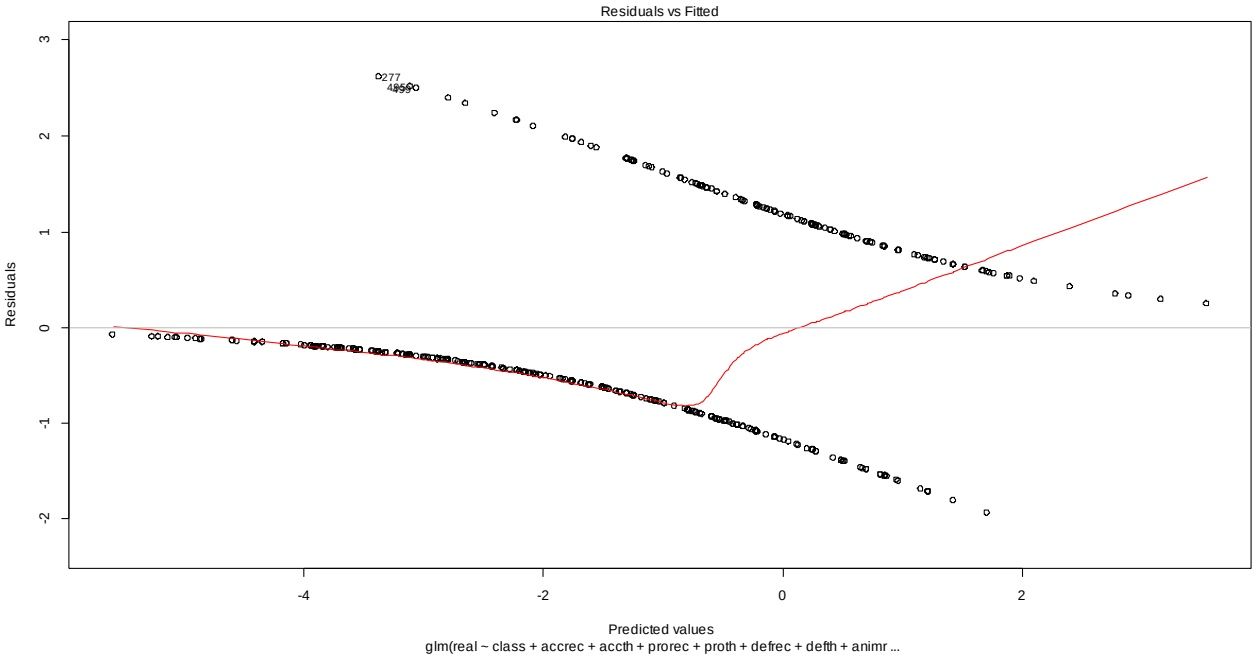
Johnson (2008) har på websidene til boken publisert de kodete dataene som lå til grunn for Bresnans første modell. Dette gjør det mulig å sammenlikne rådataene i engelsk og norsk. Bresnans data er hentet fra SWITCHBOARD-korpuset og Wall Street Journal, og inneholder et titalls ditransitive verb. Hvis man henter ut de totalt 905 treffene som stammer fra Wall Street Journal fra Bresnans datagrunnlag, kan man sammenlikne disse med de 511 treffene fra Aftenposten, som også gjelder et titalls ditransitive verb. På bakgrunn av disse rådataene kan man sette opp følgende tabell:

| **Ulike verb** | | **Wall Street Journal** | | **Aftenposten** | |
|---|---|---|---|---|---|
| faktor | verdi | NP | PP | NP | PP |
| Aksesserbarhet: recipient | given | **41,4 %** | 6,2 % | **48,5 %** | 6,1 % |
| | notgiven | 20,1 % | 32,3 % | 23,7 % | 21,7 % |
| Aksesserbarhet: theme | given | 1,8 % | 5,0 % | 6,1 % | 5,9 % |
| | notgiven | ***59,8 %*** | 33,5 % | ***66,1 %*** | 21,9 % |
| Definitthet: recipient | definite | **47,2 %** | 21,7 % | **66,3 %** | 20,0 % |
| | indef | 14,4 % | 16,8 % | 5,9 % | 7,8 % |
| Definitthet: theme | definite | 10,9 % | 13,8 % | 13,5 % | 8,2 % |
| | indef | ***50,6 %*** | 24,6 % | ***58,7 %*** | 19,6 % |
| Pronominalitet: recipient | pronom | **16,2 %** | 0,2 % | **31,5 %** | 1,4 % |
| | nonpro | 45,3 % | 38,2 % | 40,7 % | 26,4 % |
| Pronominalitet: theme | pronom | 0,4 % | 1,7 % | 0,4 % | 1,0 % |
| | nonpro | ***61,1 %*** | 36,8 % | ***71,8 %*** | 26,8 % |
| Animathet: recipient | animate | **57,8 %** | 33,3 % | **46,4 %** | 11,0 % |
| | inanim | 3,8 % | 5,2 % | 25,8 % | 16,8 % |
| Animathet: theme | animate | 0,6 % | 0,1 % | 0,4 % | 0,4 % |
| | inanim | ***61,0 %*** | 38,3 % | ***71,8 %*** | 27,4 % |
| Relativ lengde | rec>theme | 9,5 % | 24,4 % | 7,0 % | 12,7 % |
| | rec<=theme | **52,0 %** | 14,0 % | **65,2 %** | 15,1 % |
| Semantisk klasse | abstract | 43,4 % | 14,8 % | 33,1 % | 7,6 % |
| | comm | 1,8 % | 2,0 % | 8,8 % | 4,3 % |
| | fut trans | 1,2 % | 0,1 % | | |
| | prev trans | 4,9 % | 0,2 % | | |
| | transfer | 10,3 % | 21,3 % | 30,3 % | 15,9 % |

Tabell: Faktorer (engelsk versus norsk)

De **uthevede** tallene viser at for aksesserbare, definitte, pronominale og/eller animate recipienter vil recipienten i begge språk helst realiseres som NP. De ***kursivert uthevede*** tallene viser at ikke-aksesserbar, indefinitt, ikke-pronominal og/eller ikke-animat theme likeledes gir at recipienten helst realiseres som NP. Samme preferanse gjelder hvis recipienten ikke er lengre enn theme. Den eneste klare ulikheten i preferanser gjelder verb brukt i betydningen fysisk overføring (transfer): i Wall Street Journal vil to av tre slike gi PP-realisering av recipienten, i Aftenposten vil to av tre gi NP-realisering. Dette kan ha sammenheng med at mange ikke vil bruke to pronomener etter verbet i engelsk (Bresnan og Nikitina, 2003, s. 20), mens det er helt greit på norsk. Sammenlikn for eksempel *Gi meg det* kontra ??*Give me it*. På grunn av eventuelle ulikheter i kodingen av dataene, er det vanskelig å sammenlikne selve prosentandelene for hver enkelt faktor på tvers av språkene.

Følgende graf viser hvordan treffene fordeler seg i forhold til modellenes prediksjoner. Grafen viser at mange PP-er feilpredikeres som NP-er (øvre venstre kvadrant):




Graf: Aftenposten

## 4.5 Kryssvalidering av modellene

For å teste om modellene er overførbare på ukjente data, har jeg brukt talespråksmodellen på skriftspråksdata, og omvendt, hele tiden begrenset til verbet *gi*. Resultatet av denne kryssvalideringen er som følger:

Talespråksmodellen (modelT) anvendt på skriftspråksdata (numavis):

```
> table(numavis$real,predict(modelT,numavis)>0.0)
```

| | FALSE | TRUE | % korrekt predikert |
|---|---|---|---|
| NP | 1830 | 447 | 80,4 % |
| PP | 120 | 330 | 73,3 % |

Dette viser at talespråksmodellen predikerer skriftspråksdata måtelig godt. Totalt predikeres 79,2 % av treffene riktig (1830 + 330 = 2160 av 2727). Jamfør at modellen predikerer talespråksdataene den er basert på med 96,5 % treffsikkerhet, og at andelen NP-recipienter i dette skriftspråkskorpuset er 83,5 %, slik at man faktisk ville få høyere treffsikkerhet ved alltid å predikere NP enn ved å bruke denne modellen.

Skriftspråksmodellen (modelS) anvendt på talespråksdata (numtale):

```
> table(numtale$real,predict(modelS,numtale)>0.0)
      FALSE  TRUE      % korrekt predikert
  NP    163     3      98,2 %
  PP     11    23      67,6 %
```

Dette viser at skriftspråksmodellen predikerer talespråksdata en del bedre. Totalt predikeres 93,0 % av treffene riktig (163 + 23 = 186 av 200). Jamfør at modellen predikerer skriftspråksdataene den er basert på med 87,0 % treffsikkerhet, og at andelen NP-recipienter i dette talespråkskorpuset er 83,0 %, slik at man forbedrer treffsikkerheten med ti prosent-poeng ved å bruke denne modellen i stedet for alltid å predikere NP.

Det viser seg altså at de to modellene oppfører seg ulikt. Talespråksmodellen har meget god treffsikkerhet på sine egne grunnlagsdata (den gir 96,5 - 83,0 = 13,5 prosentpoeng bedre prediksjoner enn nullmodellen som alltid velger NP), men har begrenset overføringsverdi til skriftspråksdata. Skriftspråksmodellen har måtelig god treffsikkerhet på egne grunnlagsdata (den gir 87,0 - 83,5 = 3,5 prosentpoeng bedre prediksjoner enn nullmodellen), men har god overføringsverdi til talespråksdata. Siden andelen NP-recipienter i skrift og tale er tilnærmet identisk, med henholdsvis 83,5 % og 83,0 %, er det lite trolig at grunnen til denne ulikheten ligger i systematiske forskjeller mellom skrift og tale som sådan. Det er derimot enkelte ting som tyder på at disse aggregerte tallene skjuler at de underliggende korpusene er differensierte med hensyn til en formalitetsparameter. I skrift er det som tidligere nevnt i kapittel 4.1.1, en tydelig forskjell mellom abonnementsaviser og typiske ukeblad hva andelen PP-er angår, med henholdsvis 16-19 % og 6-7 %. I tale er det en tilsvarende forskjell mellom de ustrukturerte samtalene i BigBrother-korpuset (11 % PP-er) og de til dels halvformelle

samtalene og intervjuene i Nota-Oslo (25 % PP-er), selv om datagrunnlaget her er tynt. (Om karakteriseringen av Nota-Oslo som halvformelt, se websidene til korpuset, hvor opptakssituasjonene karakteriseres som ”et halvformelt intervju og en uformell samtale”.)

Bresnan et al (2007, s. 88), som riktignok ikke bare angår verbet *give*, viser tilsvarende forskjeller mellom to korpus. SWITCHBOARD-korpuset, som består av ustrukturerte telefonsamtaler, har 21 % PP-er, mens Wall Street Journal-korpuset, som består av nyheter og finansiell reportasje, har 38 % PP-er. For å kontrollere for denne forskjellen, inkluderte de en modalitets-parameter i sine endelige modeller, som i realiteten bare kodet for medium: tale eller skrift. De norske dataene ovenfor tyder på at man i stedet kan bruke en formalitets-parameter som avhenger av kilde: formell (for eksempel for abonnementsaviser og Nota-Oslo) versus uformell (for eksempel for ukeblad og Big Brother).

For å teste denne muligheten, laget jeg en kombinert modell for både skriftspråk og talespråk, med grad av formalitet som egen parameter. Den kombinerte modellen beskrives og kommenteres i neste delkapittel.

## 4.6 Kombinert modell for verbet gi

For å teste om grad av formalitet har betydning for valg av dativalternativ, laget jeg først en kombinert modell for både skrift- og talespråk, og deretter bygget jeg denne ut med en egen faktor for formalitet, der *formell* er operasjonalisert til å gjelde for tradisjonelle abonnementsaviser og det halvformelle Nota-Oslo, mens alle andre kilder, for eksempel ukeblad og Big Brother, kodes som *uformelle*.

Datamaterialet for den kombinerte modellen omfatter 2927 treff. Modellen uten egen faktor for grad av formalitet gav følgende oppsummering og parameterkoeffisienter (resultater i `Courier New`, mine kommentarer i Times New Roman):

```
> summary(modelK)

Call:
glm(formula = real ~ class + accrec + accth + prorec + proth +
    defrec + defth + animrec + animth + logdiff, family =
    binomial,data = numkombi)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3817  -0.5195  -0.2033  -0.1038   3.2333

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.23502    0.16072  -1.462  0.14367
classc       1.04993    0.22777   4.610 4.03e-06 ***
classt       1.38157    0.17950   7.697 1.40e-14 ***
accrec      -0.48504    0.17000  -2.853  0.00433 **
accth        0.89155    0.33885   2.631  0.00851 **
prorec      -1.60487    0.25312  -6.340 2.29e-10 ***
proth        2.36927    0.43344   5.466 4.60e-08 ***
defrec      -0.84023    0.16207  -5.184 2.17e-07 ***
defth       -0.03863    0.16462  -0.235  0.81448
animrec     -0.75281    0.13265  -5.675 1.38e-08 ***
animth       1.17751    0.98743   1.193  0.23306
logdiff      1.18662    0.08117  14.619  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2625.2  on 2926  degrees of freedom
Residual deviance: 1718.9  on 2915  degrees of freedom
AIC: 1742.9

Number of Fisher Scoring iterations: 6
```

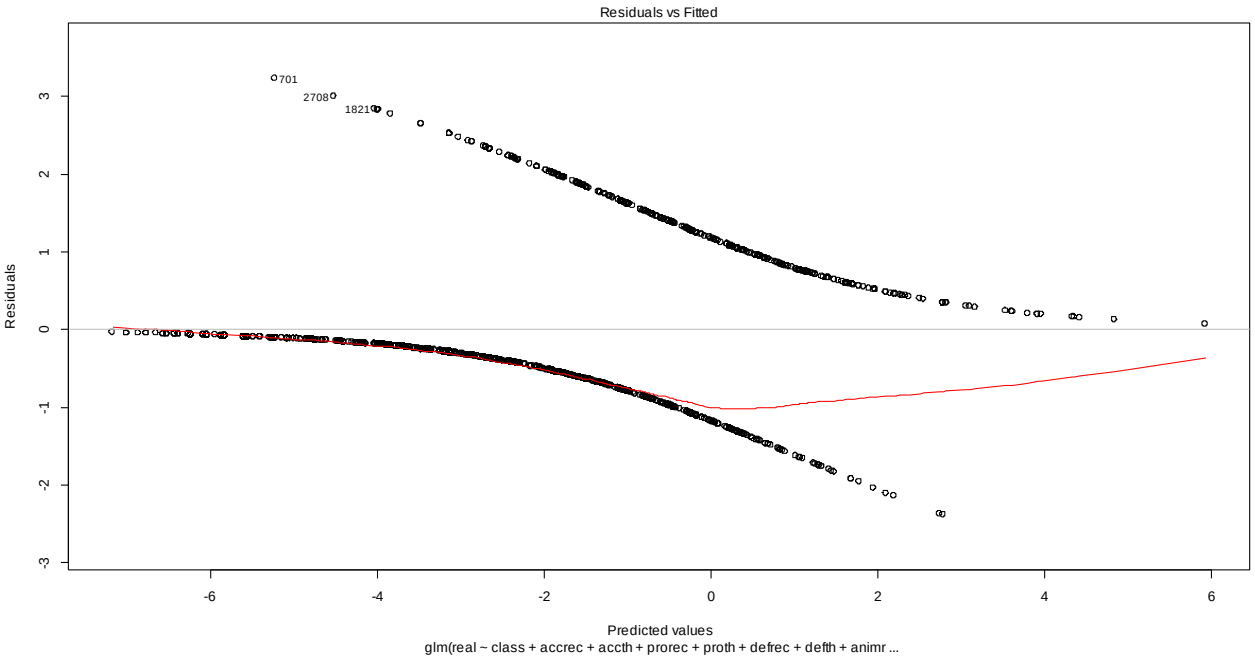
Modellen er tilnærmet lik modellen for aviser alene, og skiller seg fra den ved at også aksesserbarhet av theme (accth) er signifikant. Likheten skyldes nok at aviser utgjør hovedtyngden av datagrunnlaget (2727 av de 2927 treffene).

Modellens prediksjoner er som følger, krysstabulert mot faktisk realisering:

```
> table(numkombi$real,predict(modelK)>0.0)
```

| | FALSE | TRUE | % korrekt predikert |
|---|---|---|---|
| NP | 2350 | 93 | 96,2 % |
| PP | 277 | 207 | 42,8 % |

Følgende graf viser hvordan treffene fordeler seg i forhold til modellens prediksjoner:




Graf: Gi (kombinert)

Krysstabuleringen viser at modellen har noe bedre treffsikkerhet enn modellen for aviser alene, men det viser seg at det ikke er noen forbedring når man bryter opp tallene på henholdsvis skriftspråksdata og talespråksdata. Den tilsynelatende forbedringen totalt sett, er altså bare en effekt av aggregerte data. Også grafen viser at mange PP-er feilpredikeres som NP-er: punktene i øvre venstre kvadrant i grafen, mens punktene i nedre høyre kvadrant viser NP-er feilpredikert som PP-er.

Den kombinerte modellen kan bygges ut til en modell med en egen faktor for grad av formalitet (som funksjon av kilde), ved en enkel transformasjon gjort i R:

```
> difkombi <- transform(numkombi, formality = mod=='Ad' | mod=='Af' | mod=='Bb' | mod=='BT' | mod=='Nota' | mod=='SA' | mod=='VL')
```

Denne transformasjonen sørger for at alle treff fra kildene Adresseavisen, Aftenposten, Bondebladet, Bergens Tidende, Stavanger Aftenblad, Vårt Land og Nota-Oslo kodes som formelle, mens alle treff fra andre kilder (Dagbladet, Det Nye, Familien, Hennes, HP, KK, Motor, Universitas, VG, Big Brother og Taus) kodes som uformelle.

Datamaterialet for denne modellen, med egen faktor for grad av formalitet, omfatter altså også 2927 treff. Modellen gav følgende oppsummering og parameterkoeffisienter (resultater i `Courier New`, mine kommentarer i Times New Roman):

```
> summary(modelL)

Call:
glm(formula = real ~ class + accrec + accth + prorec + proth +
    defrec + defth + animrec + animth + logdiff + formality,
    family = binomial, data = difkombi)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4167  -0.5128  -0.2071  -0.0985   3.2040

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.53906    0.24125  -2.234  0.02546 *
classc        1.06777    0.22900   4.663 3.12e-06 ***
classt        1.38232    0.17956   7.698 1.38e-14 ***
accrec       -0.49146    0.17026  -2.886  0.00390 **
accth         0.89008    0.33959   2.621  0.00877 **
prorec       -1.54798    0.25482  -6.075 1.24e-09 ***
proth         2.46165    0.43581   5.648 1.62e-08 ***
defrec       -0.84495    0.16229  -5.206 1.93e-07 ***
defth        -0.02490    0.16479  -0.151  0.87992
animrec      -0.74295    0.13276  -5.596 2.19e-08 ***
animth        1.13124    0.99374   1.138  0.25497
logdiff       1.17904    0.08125  14.512  < 2e-16 ***
formalityTRUE  0.33491    0.19758   1.695  0.09006 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2625.2  on 2926  degrees of freedom
Residual deviance: 1715.9  on 2914  degrees of freedom
AIC: 1741.9

Number of Fisher Scoring iterations: 6
```

Modellen er tilnærmet lik den kombinerte modellen uten egen formalitets-parameter. Faktoren formalitet (formality) er ikke signifikant på 0,05-nivået, men så vidt signifikant på 0,1-nivået. For å undersøke om den likevel kan bidra til å bedre treffsikkerheten til den kombinerte modellen, er det nødvendig å se på modellens prediksjoner, som er som følger, krysstabulert mot faktisk realisering:

```
> table(difkombi$real,predict(modelL)>0.0)
      FALSE    TRUE       % korrekt predikert
  NP  2349       94       96,2 %
  PP   278      206       42,6 %
```

Modellen viser ingen forbedring av treffsikkerheten. Man kan altså konkludere med at grad av formalitet (som funksjon av kilde) ikke synes å påvirke valg av dativalternativ i skrift og tale. Den påtakelige forskjellen i andel PP-er i ulike media må altså tilbakeføres til andre faktorer enn grad av formalitet.

Forskjellen kan illustreres av følgende krysstabulering av faktisk realisering kontra grad av formalitet som funksjon av kilde:

```
> table(difkombi$real,difkombi$formality)
     FALSE TRUE
  NP   528 1915
  PP    52  432
```

Krysstabuleringen viser at de kildene som er kodet som uformelle, har en gjennomsnittelig PP-andel på 9,0 %, mens de kildene som er kodet som formelle, har en gjennomsnittelig PP-andel på 18,4 %, altså det dobbelte av de uformelle.

## 4.7 Interaksjonseffekter

Alle foregående modeller bygger på at parametrene (faktorene) er uavhengige, og at eventuelle interaksjonseffekter mellom dem ikke i vesentlig grad påvirker modellens

prediksjoner. Dette er faktisk en av styrkene ved logistisk regresjon, men metoden tillater også at man bygger interaksjonseffekter inn i modellen. I statistikkverktøyet R kan dette gjøres ved i modellformelen å angi at man vil se både på hver faktor for seg, og alle mulige kombinasjoner av utvalgte faktorer. Det er da nærliggende å se på interaksjoner mellom aksesserbarhet, definitthet, pronominalitet og animathet for recipient og theme hver for seg. Jeg bygget derfor ut den kombinerte modellen for *gi* i skrift og tale med disse interaksjonseffektene. Resultatet ble som følger (resultater i `Courier New`, mine kommentarer i Times New Roman):

```
> summary(modelI)

Call:
glm(formula = real ~ class + accrec * prorec * defrec *
animrec + accth * proth * defth * animth + logdiff, family =
binomial, data = numkombi)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6486  -0.4898  -0.1863  -0.0748   3.4297

Coefficients: (8 not defined because of singularities)
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.13392    0.22886  -0.585  0.55843
classc                1.08882    0.23116   4.710 2.47e-06 ***
classt                1.34995    0.18341   7.360 1.84e-13 ***
accrec                1.30703    1.16016   1.127  0.25991
prorec                0.38887    2.11375   0.184  0.85404
defrec               -1.10024    0.25370  -4.337 1.45e-05 ***
animrec              -0.92080    0.29237  -3.149  0.00164 **
accth                 1.36239    0.67707   2.012  0.04420 *
proth                 1.97848    1.24971   1.583  0.11339
defth                -0.06286    0.17281  -0.364  0.71602
animth              -10.65099  324.22212  -0.033  0.97379
logdiff               1.19136    0.08281  14.387  < 2e-16 ***
accrec:prorec        -0.11046    1.64971  -0.067  0.94662
accrec:defrec        -1.39981    1.18406  -1.182  0.23712
prorec:defrec        -0.89700    1.41494  -0.634  0.52611
accrec:animrec       -4.87225    1.69118  -2.881  0.00396 **
prorec:animrec        0.62112    1.68874   0.368  0.71302
defrec:animrec        0.43744    0.33727   1.297  0.19463
accth:proth           0.60359    1.01373   0.595  0.55156
```

```
accth:defth                  -0.20487    0.79804  -0.257  0.79739
proth:defth                   0.16799    1.43746   0.117  0.90697
accth:animth                -15.62348  882.74492  -0.018  0.98588
proth:animth                 23.52997 1248.38794   0.019  0.98496
defth:animth                 13.83342  324.22604   0.043  0.96597
accrec:prorec:defrec               NA         NA      NA       NA
accrec:prorec:animrec        -2.34400    1.80031  -1.302  0.19292
accrec:defrec:animrec         4.45439    1.72328   2.585  0.00974 **
prorec:defrec:animrec              NA         NA      NA       NA
accth:proth:defth                  NA         NA      NA       NA
accth:proth:animth                 NA         NA      NA       NA
accth:defth:animth                 NA         NA      NA       NA
proth:defth:animth                 NA         NA      NA       NA
accrec:prorec:defrec:animrec       NA         NA      NA       NA
accth:proth:defth:animth           NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2625.2  on 2926  degrees of freedom
Residual deviance: 1674.1  on 2901  degrees of freedom
AIC: 1726.1

Number of Fisher Scoring iterations: 13
```

Enkeltfaktorene klasse (class), definitthet av recipient (defrec), animathet av recipient (animrec), aksesserbarhet av theme (accth) og relativ lengde (logdiff) er fortsatt signifikante i denne modellen. Fortegnene til disse faktorenes modellkoeffisienter (kolonnen ’Estimate’ til venstre) viser dessuten at de trekker i samme retning som i en modell uten interaksjonseffekter: definitt og/eller animat recipient tenderer mot å NP-realiseres (negative verdier), aksesserbar theme og recipient som er lengre enn theme tenderer mot at recipienten realiseres som en PP (positive verdier). Dessuten er kombinasjonen aksesserbarhet og animathet av recipient (accrec:animrec) og den tresidige kombinasjonen aksesserbarhet, definitthet og animathet av recipient (accrec:defrec:animrec) også signifikante, men trekker i hver sin retning. Det mest påfallende er likevel at pronominalitet av recipient ikke lenger er signifikant verken som enkeltfaktor eller som deltaker i en interaksjon mellom faktorer. Dette

bekrefter imidlertid det noe uklare mønsteret for pronominalitet av recipient fra tabuleringen av rådataene i kapittel 4.1.2.

Modellen har ikke blitt vesentlig forbedret hva 'deviance' angår. Det er altså fortsatt en betydelig restvarianse i datamaterialet som ikke kan forklares verken av modellparametrene eller interaksjonseffekter mellom dem.

For å undersøke om interaksjonseffektene likevel kan bidra til å bedre treffsikkerheten til den kombinerte modellen, er det igjen nødvendig å se på modellens prediksjoner, som er som følger, krysstabulert mot faktisk realisering:

```
> table(numkombi$real,predict(modelI)>0.0)
```

| | FALSE | TRUE | % korrekt predikert |
|---|---|---|---|
| NP | 2352 | 91 | 96,3 % |
| PP | 270 | 214 | 44,2 % |

Modellen er bare marginalt bedre enn en modell uten interaksjonseffekter. Siden slike interaksjonseffekter kompliserer modellformelen, og spesielt gjør det vanskeligere å tolke parameterkoeffisientene, er den lille forbedringen i prediksjonskraft neppe nok til at man bør foretrekke en modell med interaksjonseffekter foran en uten.

Et forbehold må tas om at andre interaksjoner enn de jeg her har tatt for meg kan spille en rolle. I engelsk har Bresnan og Hay (2008, s. 251) for eksempel vist at interaksjonen mellom aksesserbarhet av theme og semantisk klasse er signifikant. Spesifikt viser de at ikke-aksesserbart theme og *gi* brukt i betydningen fysisk overføring øker sannsynligheten for en dobbelt-objekt-konstruksjon betydelig. I prinsippet er det ikke noe i veien for å undersøke alle logisk mulige kombinasjoner av faktorer i en modellformel i R, men resultatet av en slik øvelse blir noe vanskelig å overskue.

## 4.8 En mixed-effects modell for Aftenposten

Alle modellene som hittil er beskrevet, er eksempler på vanlig logistisk regresjon. Faktoren verbbruk, som hittil ikke er benyttet, er dårlig tilpasset standard logistisk regresjon, siden den deler datasettet opp i et relativt stort antall disjunkte grupper. Dataene fra Aftenposten inneholder for eksempel totalt 31 ulike verbbetydninger (verbbruksmåter) fordelt på inntil tre ulike bruksmåter for hvert av 20 ulike verb, hvorav *gi* i abstrakt, kommunikativ og overførings-bruk alene utgjør tre bruksmåter. For å kunne inkludere også denne faktoren, kan man gå over til en såkalt mixed effect logistisk regresjonsmodell. Slike modeller inneholder to komponenter: standard logistisk regresjon for de faste (verbuavhengige) faktorene og en antatt normalfordelt randomisert verbbruksparameter. Modellformelen for en slik mixed effect modell blir da: $f(z) = 1/(1+e^{-z})$ der $z = X_i\beta + \mu_i$ og $\mu_i$ er normalfordelt. I statistikkprogrammet R er mixed effect logistisk regresjon ikke en del av standardoppsettet, men tilgjengelig via kommandoen `library(MASS)` og **glmmPQL**-prosedyren. Følgende utskrift er resultatet av å bruke en slik mixed effect modell på de 511 treffene for ulike ditransitive verb i Aftenposten (resultater i `Courier New`, mine kommentarer i Times New Roman):

```
> summary(modelM)
Linear mixed-effects model fit by maximum likelihood
 Data: numaften
  AIC BIC logLik
   NA  NA     NA

Random effects:
 Formula: ~1 | vsense
        (Intercept)  Residual
StdDev:    1.599145 0.9645673

Variance function:
 Structure: fixed weights
 Formula: ~invwt
Fixed effects: real ~ -1 + class + accrec + accth + prorec +
proth + defrec + defth + animrec + animth + logdiff
             Value Std.Error  DF   t-value p-value
classa   1.9273547 0.8506159  28  2.265834  0.0314
```

```
classc   1.8447456 1.1568598  28  1.594615  0.1220
classt   1.9183167 0.6461686  28  2.968756  0.0061
accrec  -0.4590931 0.3377346 472 -1.359331  0.1747
accth    0.4249807 0.4474430 472  0.949799  0.3427
prorec  -1.7024181 0.5441624 472 -3.128511  0.0019
proth    0.6971705 1.2760417 472  0.546354  0.5851
defrec  -0.9376946 0.3675245 472 -2.551380  0.0110
defth    0.1222864 0.3701844 472  0.330339  0.7413
animrec -0.6557345 0.2992287 472 -2.191416  0.0289
animth  -0.7884966 2.5881206 472 -0.304660  0.7608
logdiff  0.8890799 0.1684773 472  5.277149  0.0000

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-2.3031963 -0.4087366 -0.1690970  0.1808709  9.2207169

Number of Observations: 511
Number of Groups: 31
```

Modellutskriften stjernekoder ikke signifikans for parametrene, men kolonnen ’p-value’ helt til høyre viser at pronominalitet av recipient (prorec), definitthet av recipient (defrec), animathet av recipient (animrec) og relativ lengde av recipient og theme (logdiff) alle er signifikante med $p < 0,05$. Det samme er bruksmåtene abstrakt (classa) og fysisk overføring (classt). Fortegnene til de ulike faktorene (kolonnen ’Value’ til venstre) er de samme som for den ordinære logistiske modellen i kapittel 4.4, med unntak av animathet av theme (animth), som uansett er insignifikant i begge disse modellene. Aksesserbar, pronominal, definitt og/eller animat recipient trekker altså i retning av at recipienten realiseres som en NP (negative verdier), mens aksesserbar, pronominal og/eller definitt theme trekker i retning av at recipienten realiseres som en PP (positive verdier). Relativ fraselengde (logdiff) viser som før at hvis recipienten er lengre enn theme, trekker dette i retning av at recipienten PP-realiseres (positiv verdi).

For å undersøke om denne mixed effect modellen har bedre treffsikkerhet enn den ordinære modellen, er det som vanlig nødvendig å se på modellens prediksjoner, som er som følger, krysstabulert mot faktisk realisering:

```
> table(numaften$real,predict(modelM)>0.0)
```

| | FALSE | TRUE | % korrekt predikert |
|---|---|---|---|
| NP | 350 | 19 | 94,9 % |
| PP | 41 | 101 | 71,1 % |

Modellen er betydelig bedre enn den ordinære modellen. Totalt predikeres 88,3 % av treffene korrekt, mot 80,0 % for den ordinære modellen i kapittel 4.4. En nærmere kryssvalidering av denne modellen mot taledataene og skriftdataene for verbet *gi* alene, viser imidlertid ingen positiv effekt på treffsikkerheten. Modellen tilsvarer Bresnan et als (2007, s. 86) modell B (for engelsk tale), som er oppgitt å ha en treffsikkerhet på 95 %.

Kapittel 4.7 viste at det også kan forekomme interaksjonseffekter mellom modellfaktorene. En måte å finne ut hvilke interaksjonseffekter man eventuelt kan supplere modellen med, er å se på en krysstabulering av korrelasjonene mellom koeffisientverdiene for modellparametrene. I dette tilfellet gir en slik krysstabulering følgende tabell:

```
Correlation:
        classa classc classt accrec accth  prorec proth  defrec defth  animrc animth
classc   0.128
classt   0.235  0.172
accrec  -0.025  0.009  0.003
accth   -0.046 -0.020  0.018 -0.165
prorec   0.012 -0.033 -0.002 -0.413  0.109
proth   -0.002 -0.040  0.036 -0.014 -0.076 -0.013
defrec  -0.318 -0.246 -0.408 -0.215  0.020 -0.002 -0.046
defth   -0.054  0.001 -0.132  0.105 -0.347 -0.021 -0.074 -0.183
animrec -0.214 -0.131 -0.328 -0.013 -0.046 -0.266 -0.025  0.079  0.076
animth  -0.037 -0.006 -0.056 -0.048  0.010  0.071 -0.116  0.031 -0.017 -0.022
logdiff  0.001  0.095  0.043  0.099 -0.042  0.034 -0.090  0.159 -0.009 -0.093 -0.027
```

Dette viser at det er en klar negativ korrelasjon mellom koeffisientverdiene for aksesserbarhet (accrec) og pronominalitet (prorec) av recipienten og en litt mindre negativ korrelasjon mellom koeffisientverdiene for aksesserbarhet (accrec) og definitthet (defrec) av recipienten. Det er også en forholdsvis klar negativ korrelasjon mellom koeffisientverdiene for pronominalitet (prorec) og animathet (animrec) av recipienten. Hva theme angår, er det likeledes en klar negativ korrelasjon mellom koeffisientverdiene for aksesserbarhet (accth) og definitthet (defth). Mer interessant er det kanskje at det også er klare negative korrelasjoner mellom koeffisientverdiene for abstrakt verbbruk (classa) respektive overføringsbruk (classt)

og både definitthet (defrec) og animathet (animrec) av recipienten. Alle disse korrelasjonene bekrefter at det kan være formålstjenlig å undersøke interaksjonseffekter mellom utvalgte faktorer, og peker på noen kombinasjoner av faktorer som det kan være interessant å inkludere i modellen. Negativ korrelasjon mellom koeffisientverdier tyder på positiv korrelasjon mellom angjeldende faktorer, slik at de tenderer til å forsterke hverandre. Siden kapittel 4.7 også viser at det kan være vanskelig å tolke modeller med interaksjonseffekter, avstår jeg fra å forfølge dette her.

# 5 Drøfting

Av modellresultatene i foregående kapittel er det mulig å trekke følgende konklusjoner:

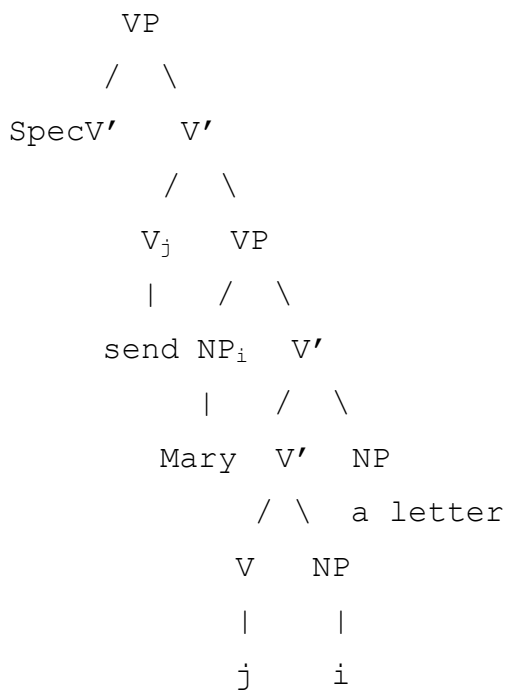
- Verbet *gi* er spesielt, for så vidt som det er vanskeligere å predikere dativalternering for dette verbet, enn for andre ditransitive verb.
- Det er generelt lettere å korrekt predikere V-NP-NP enn V-NP-PP.
- Det er generelt lettere å korrekt predikere taledata enn skriftdata.
- Aksesserbarhet, pronominalitet, definitthet, animathet og fraselengde har betydning for valg av konstruksjonstype, men det har også bruksmåten (abstrakt, kommunikativ eller fysisk overføring) og verbets mer spesifikke betydning (for eksempel verbet *gi* i abstrakt betydning).

Verbet *gi* kan oppfattes som et naturlig laboratorium for å undersøke hvilke andre faktorer som kan tenkes å påvirke dativalterneringen, for så vidt som dette verbet synes å være mest variabelt. Man kan antakelig gå ut fra at marginale effekter av andre faktorer enn de jeg har undersøkt, vil gi større effekter for valg av konstruksjonstype for dette verbet enn for noen andre ditransitive verb.

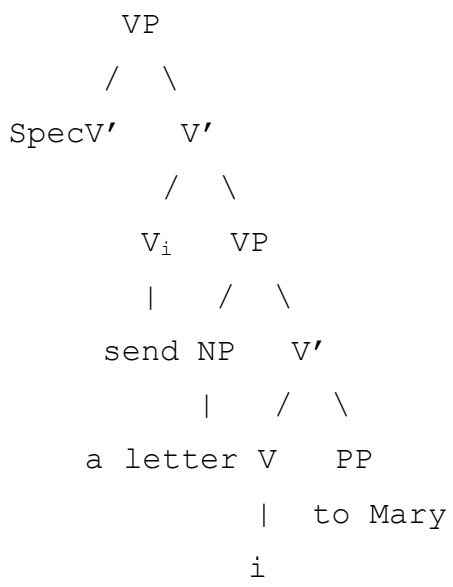
Hva de faktisk undersøkte faktorene angår, underbygger både rådataene og de ulike matematiske modellene for valg av dativalternativ den grunnleggende forventningen om at denne vekslingen påvirkes av disse faktorene. Recipienter som er aksesserbare i konteksten tenderer til å realiseres som NP-er rett etter verbet i en dobbelt-objekt-konstruksjon V-NP-NP. Det samme gjør recipienter som er definitte, pronominale og/eller animate. Det vil si at recipienter som er prominente på minst en av disse prominensskalaene, tenderer til å realiseres i en prominent posisjon i setningen. På samme måte tenderer theme som er prominent på minst en av disse skalaene til å realiseres i en prominent posisjon, som direkte objekt rett etter verbet i en preposisjonsfrase-konstruksjon V-NP-PP.

Hvis man tenker seg dobbelt-objekt-konstruksjonen V-NP-NP som et eksempel på et såkalt Larsonsk skall, vil recipienten c-kommandere theme, og i denne forstand være mest prominent. For å knytte an til semantiske forklaringer på dativalterneringen, kunne man kanskje se på denne formen for syntaktisk prominens som en syntaktisk realisering av det

semantiske forholdet mellom possessor og possessert ting. Dette blir da som i Larson (1988, s. 354) som gir følgende syntaktiske tre for *send* med dobbelt objekt (lett tilpasset):




I preposisjonsfrase-konstruksjonen V-NP-PP gir et tilsvarende tre at preposisjonsfrasen (og dermed recipienten) blir c-kommandert av theme. Dermed kan man kanskje på tilsvarende måte oppfatte denne konstruksjonen som en syntaktisk realisering av det semantiske forholdet mellom en ting og tingens lokasjon. Larson (1988, s. 343) gir følgende syntaktiske tre for *send* med preposisjonsfrase (lett tilpasset):



I en ikke-transformasjonell teori som LFG er det tilsvarende et naturlig prominenshierarki mellom de fire grammatiske hovedfunksjonene:

$$SUBJ > OBJ > OBJ_{\theta} > OBL_{\theta}$$

De to siste har trekket [+restricted] i LFGs Lexical Mapping Theory (LMT) og kan dermed bare ha den spesielle tematiske rollen θ (for eksempel recipient), mens de to første kan ha flere ulike tematiske roller (Lødrup, 2011, s. 146). Dette hierarkiet gjenspeiles i syntaksen for dativalterneringen for så vidt som både dobbelt-objekt- og preposisjonsfrase-konstruksjonen realiserer OBJ i den syntaktisk prominente postverbale posisjonen foran henholdsvis $OBJ_{THEME}$ og $OBL_{RECIPIENT}$, som vist i kapittel 2. (Igjen forutsetter dette at man ikke benytter den tradisjonelle analysen av doble objekter i norsk som indirekte og direkte objekt.)

Bresnan og Ford (2010, s. 183) gjengir et diagram, som direkte viser sammenhengen mellom ulike prominenshierarkier og syntaktisk realisering, hva de kaller et kvalitativt bilde av den kvantitative formen for 'harmonic alignment' som blant andre modellene i foregående kapittel representer. En tilpasning av dette diagrammet gir følgende figur for preferert realisering:

| | | | |
|---|---|---|---|
| | aksesserbar | > | **ikke-aksesserbar** |
| | animat | > | **ikke-animat** |
| | definitt | > | **ikke-definitt** |
| | pronominal | > | **ikke-pronominal** |
| | kort | > | **lang** |
| V | $NP_{recipient}$ | | $\mathbf{NP_{theme}}$ |
| V | $NP_{theme}$ | | $\mathbf{PP_{recipient}}$ |

Diagrammet skal forstås som at recipienter og theme med gitte egenskaper (uthevede eller ikke uthevede) vil tendere mot å bli realisert i tilsvarende (uthevet eller ikke uthevet) posisjon i setningen. Igjen ser vi at prominente egenskaper gir prominent (postverbal) plassering.

Ved å tenke syntaktisk prominens på denne måten, kan man kanskje bygge bro over meningsforskjellene med hensyn til hvorvidt dativalternering skal forklares ut fra et monosemi-, polysemi- eller informasjonsstruktur-basert syn. Hver av disse

forklaringsmodellene har sine anvendelser, men de kan knyttes sammen ved hjelp av formelle prominensrelasjoner.

Selv om rådataene og de matematiske modellene som er presentert i forrige kapittel understøtter dette synet på dativalternering, kan de ikke forklare all variasjon i faktisk språkbruk. Enten må det finnes flere faktorer som påvirker alterneringen, eller også er det i siste instans et betydelig innslag av tilfeldig veksling: alternative konstruksjoner i streng forstand. Bresnan et al (2007) har, som tidligere nevnt, påvist at noen andre faktorer, som for eksempel person og tall for recipient og theme, eller priming gjennom parallelle strukturer i to påfølgende setninger, har innvirkning på valg av alternativ. Bresnan et al finner også at modalitet (skrift eller tale) har betydning, mens jeg ikke finner at grad av formalitet har statistisk signifikant innvirkning på valget. Uansett vil det være en rest-variasjon som de matematiske modellene ikke kan forklare. I mine modeller gjelder dette spesielt for preposisjonsfrasekonstruksjoner V-NP-PP, som modellene ofte predikerer til å skulle være dobbelt-objekt-konstruksjoner V-NP-NP i stedet. Man kan i den forbindelse spekulere på om det kan være en frekvens-effekt her. Siden V-NP-NP-konstruksjonene står for om lag 83 % av tilfellene (for verbet *gi*), kan man tenke seg at språkbrukerne har bygget sterkere intuisjoner om disse enn om V-NP-PP-konstruksjonen. I så fall burde det være mindre variasjon (avvik fra modellene) for dobbelt-objekt-konstruksjonen enn for preposisjonsfrase-konstruksjonen, noe grafene og krysstabuleringene i forrige kapittel faktisk viser. Selv om disse avvikene er mindre i Bresnan et als modeller for engelsk, viser de samme tendens: V-NP-NP-konstruksjoner predikeres med større treffsikkerhet enn V-NP-PP-konstruksjoner.

Nytteverdien av slike modeller kan dermed diskuteres. De kan, som alle probabilistiske modeller, bare si noe om tendenser i et gitt datamateriale, og forventninger til hva man vil finne i andre data av samme type. Prediksjonene, slik de er brukt i krysstabuleringene, er kategoriske, og reflekterer dermed dårlig den underliggende sannsynlighetsberegningen for en gitt setning. For recipienter med sannsynlighet nær 0,5 for å realiseres som PP-er, er det ikke veldig informativt å få en kategorisk prediksjon for det ene eller andre valget. De aggregerte prediksjonene for et stort datamateriale er likevel det beste kriteriet for hvor gode modellene er, spesielt i forhold til andre modeller som bygger på samme metode. Min beste modell, modellen for verbet *gi* i norsk tale, kan predikere 98,2 % av V-NP-NP-konstruksjonene, og 88,2 % av V-NP-PP-konstruksjonene. Bresnan et als (2007, s. 79) første modell for flere ulike ditransitive verb i engelsk tale, kan predikere 97 % av NP-ene og 77 % av PP-ene. Dette må

imidlertid kontrasteres ved at min kombinerte modell for verbet *gi* i skrift og tale predikerer 96,2 % av NP-ene, men bare 42,8 % av PP-ene. Bresnan et als (2007, s. 89) kombinerte modell for engelsk tale og skrift kan predikere 96 % av NP-ene og 86 % av PP-ene. Mens mine modeller blir dårligere når skrift og tale kombineres, blir Bresnan et als bedre, og dette er det vanskelig å forklare. Uansett må forklaringen søkes i skrift-korpuset og dets koding.

# 6 Oppsummering

Jeg håper å ha vist at logistisk regresjon er en interessant måte å få fram sammenhenger og mønstre i et datamateriale hentet fra ulike språkkorpus, og at matematiske modeller med fordel kan brukes når slike sammenhenger skal beskrives. På denne måten kan man forhåpentligvis komme videre fra en kvalitativ analyse bygget på enkeltlingvisters mer eller mindre pålitelige intuisjoner om hva som er grammatisk eller ugrammatisk i et gitt språk, til en empirisk fundert kvantitativ beskrivelse og analyse av hvilke konstruksjoner som faktisk brukes i språket, og med hvilken frekvens. Mange lingvister vil sikkert se interessante koblinger til diskusjonen om kompetanse versus performanse, I-språk versus E-språk, *langue* kontra *parole*, etc, i en slik tilnærming, men selv ville jeg foretrekke at man fokuserer på konkrete undersøkelser som har en sjanse til å komme til definitive resultater som alle kan enes om, uansett lingvistisk ståsted og foretrukket rammeverk. Hvis jeg likevel skulle bruke denne undersøkelsen til å si noe om hvorvidt den kan belyse skillet kompetanse versus performanse, ville jeg peke på at modellene predikerer taledata bedre enn skriftdata. Hvis man antar at skriftlige ytringer i liten grad lider av de begrensningene i performanse som ofte nevnes i forbindelse med talespråk (forsnakkelser, begrenset hukommelse, temporær kognitiv svekkelse, etc) skulle man antakelig forvente at skriftlige ytringer reflekterer underliggende regulariteter i språket (kompetansen) bedre enn talte ytringer. Denne undersøkelsen viser tvert imot at skriftlige ytringer er mer variable enn talte ytringer, for så vidt som en større andel av preposisjonsfrasekonstruksjonene da feilpredikeres av modellene. Dette reflekterer kanskje at kompetansen (grammatikken) ikke er kategorisk, men derimot bruker parametriserte regler av den typen Labov (1972, kapittel 8) i sin tid introduserte, og at disse reglene omfatter noen parametre (faktorer) som modellene i denne undersøkelsen ikke har greid å isolere, og som er spesielt virksomme når man ytrer seg i skrift. Uansett heller jeg i retning av at disse modellene ikke bare reflekterer språkbruk (performansen), men at de tyder på at grammatikken (kompetansen) også må ha en probabilistisk struktur.

Hvis jeg skulle gjøre en liknende undersøkelse om dativalternering eller andre slags alternative konstruksjoner på nytt, ville jeg ha gjort noen endringer i metoden.

For det første ville jeg ikke ha påbegynt undersøkelsen med mindre jeg hadde tilgang til et korpus som var tagget på formålstjenlig måte. I mitt tilfelle ville dette ha vært en trebank med

setninger som var ferdig analysert syntaktisk, slik at jeg lett kunne hente ut relevante treff, og slapp å sortere bort alle ikke-relevante treff manuelt. Det ville spart veldig mye tid.

For det andre ville jeg ikke ha kodet de utvalgte faktorene binært, men beholdt de underliggende prominensskalaene i kodet form. Dermed kunne man eksperimentere med ulike bruddpunkter, for eksempel for definitthet, under selve modelleringen, og foreta en mer induktivt basert analyse av hvordan prominensskalaene faktisk blir brukt. Dette ville ikke nødvendigvis ha tatt særlig mye mer tid under kodingen enn den binære metoden, og ville ha forhindret at relevante data gikk tapt under selve kodingsprosessen.

For det tredje ville jeg ha inkludert flere mulige faktorer i kodingen, for eksempel person og tall, og heller fjernet insignifikante faktorer etter modelleringen. I prinsippet burde alle faktorer som har vært foreslått i tidligere litteratur om emnet, tas med i undersøkelsen.

For det fjerde ville jeg ha sett nærmere på effekten av andre interaksjoner mellom to eller flere faktorer enn de jeg har sett på, noe som altså enkelt kan gjøres i et program som R. På denne måten kunne man undersøke både effekten av hver faktor for seg, og ta underliggende korrelasjoner mellom faktorene i betraktning. Slike interaksjonseffekter er lite undersøkt fra før, antakelig fordi man ikke har hatt de nødvendige statistiske metodene. Logistisk regresjon egner seg meget godt til slike undersøkelser.

For det femte ville jeg helst ha gjort undersøkelsen på et tospråklig korpus, slik at man ikke bare fikk en beskrivelse av sammenhenger og mønster i ett gitt språk, men kunne sammenlikne og kontrastere modeller på tvers av de to språkene. Siden man da kunne være sikker på at samme kodingsprosedyre og –kriterier var benyttet på begge språkene, ville man unngå den konstante tvilen om ulike modelleringsresultater i to språk skyldes ulikheter i språkene selv, eller bare ulikheter i (anvendelse av) metode.

For det sjette ville jeg ha sett på flere verb enn *gi*, og i større bredde. Det er mye som tyder på at verbet *gi* har litt andre egenskaper enn andre ditransitive verb, noe som riktignok gjør det velegnet til å undersøke faktorer som har mindre, men likevel signifikant, innvirkning på valg av dativalternativ, men som gjør at modeller som i hovedsak bygger på dette verbet, har relativt liten overføringsverdi til datasett med stort innslag av andre verb. Dessuten kan det tenkes at en grundigere undersøkelse av andre verb ville bringe andre faktorer på banen, eksempelvis faktorer som i større grad var knyttet til det enkelte verbs spesifikke betydning.

For det sjuende ville jeg i større grad ha benyttet mixed effects modeller av den typen som er vist i kapittel 4.8. Slike modeller tar hensyn til både faste faktorer som gjelder alle verb, og betydningen av det enkelte verbs ulike bruksmåter. Slik sett kan man fange opp ikke bare de effektene som kommer av hensyn til informasjonsstruktur og den enkelte konstituents relative kompleksitet, men også de spesifikke preferanser språkbrukerne har i forbindelse med hvert enkelt verb.

Til sist ville jeg ha bedt om hjelp til kodingen av datamaterialet. Logistisk regresjon krever mye data, og det tar tilsvarende mye tid å kode, i praksis flere månedsverk. Modellene blir ikke bedre enn kvaliteten på de dataene de bygger på, så kodingen burde også verifiseres, for eksempel ved at to personer koder samme datasett parallelt, og diskuterer tvilstilfeller seg i mellom.

# Litteraturliste


Aissen, J. 2003, "Differential object marking: iconicity vs. economy", *Natural Language & Linguistic Theory*, 21, s. 435-483.

Arnold, J., Wasow, T., Losongco, A. og Ginstrom, R. 2000, "Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering", *Language*, 76, s. 28-55.

Barðdal, J., Kristoffersen, K.E. og Sveen, A. 2011, "West Scandinavian ditransitives as a family of constructions: With a special attention to the Norwegian 'V-REFL-NP' construction", *Linguistics*, 49, s. 53-104.

Bresnan, J. 2007, "Is syntactic knowledge probabilistic? Experiments with the English dative alternation", i Featherston, S. og Sternefeld, W. (red.), *Roots: Linguistics in Search of Its Evidential Base*, Mouton de Gruyter, Berlin.

Bresnan, J., Cueni, A., Nikitina, T. og Baayen, R.H. 2007, "Predicting the dative alternation", i Bouma, G., Kraemer, I. og Zwarts, J. (red.): *Cognitive Foundations of Interpretation*, Royal Netherlands Academy of Science, Amsterdam.

Bresnan, J. og Aissen, J. 2002, "Optimality and functionality: Objections and refutations", *Natural Language & Linguistic Theory*, 20, s. 81-95.

Bresnan, J. og Ford, M. 2010, "Predicting syntax: Processing dative constructions in American and Australian varieties of English", *Language*, 86, s. 168-213.

Bresnan, J. og Hay, J. 2008, "Gradient grammar: An effect of animacy on the syntax of give in New Zealand and American English", *Lingua*, 118, s. 245-259.

Bresnan, J. og Nikitina, T. 2003, "On the Gradience of the Dative alternation", On-line at http://www.stanford.edu/~bresnan//new-dative.pdf [25.04.2012]

Bruening, B. 2010, "Double Object Constructions Disguised as Prepositional Datives", *Linguistic Inquiery*, 41, s. 287-305.

Butt, M., Dalrymple, M. og Frank, A. 1997, "An architecture for linking theory in LFG", *Proceedings of the LFG97 Conference*, University of California, San Diego.

Colleman, T. 2009, "Verb disposition in argument structure alternation: a corpus study of the dative alternation in Dutch", *Language Sciences*, 31, s. 593-611.

Dalgaard, P. 2008, *Introductory statistics with R*, Second edition, Springer.

Faarlund, J.T., Lie, S. og Vannebo, K.I. 2006, *Norsk referansegrammatikk*, Fjerde opplag, Universitetsforlaget, Oslo.

Haspelmath, M. 2005, "Ditransitive constructions: the verb 'give'", i Haspelmath, M. (red), *The World atlas of language structures*, Oxford University Press, Oxford.

Johnson, K. 2008, *Qualitative methods in linguistics*, Blackwell Publishing, Oxford.

Kittilä, S. 2006, "The anomaly of the verb 'give' explained by its high (formal and semantic) transitivity", *Linguistics*, 44, s. 569-612.

Krifka, M. 2003, "Semantic and Pragmatic Conditions for the Dative Alternation", On-line at http://amor.cms.hu-berlin.de/~h2816i3x/Publications/DativeAlternationKorea.pdf [25.04.2012]

Labov, W. 1972, *Sociolinguistic patterns*, University of Pennsylvania Press.

Larson, R.K. 1988, "On the double object construction", *Linguistic Inquiry*, 19, s. 335-391.

Levin, B. og Rappaport Hovav, M. 2002, "What alternates in the Dative Alternation?", On-line at http://grammar.ucsd.edu/courses/lign270/Levin_Hovav_2002.pdf [25.04.2012]

Levin, B. og Rappaport Hovav, M. 2005, *Argument Realization*, Cambridge University Press, Cambridge.

Lødrup, H. 1995, "The realization of benefactives in Norwegian", *Papers from the 31. Regional meeting of the Chicago linguistic society*, s. 317-328.

Lødrup, H. 2011, "Lexical-Functional Grammar: Functional structure", i Borsley, R.D. og Börjars, K. (red.), *Non-transformational syntax: Formal and explicit models of grammar*, Blackwell Publishing, Oxford.

Manning, C.D. 2003, "Probabilistic syntax", i Bod, R., Hay, J. og Jannedy, S. (red.) 2003, *Probabilistic linguistics*, MIT Press, Cambridge.

Newman, J. (red.) 1998, *The linguistics of giving*, John Benjamins Publishing Company, Amsterdam.

Newmeyer, F.J. 2002, "Optimality and functionality: a critique of functionally-based optimality-theoretic syntax", *Natural Language & Syntactic Theory*, 20, s. 43-80.

Pinker, S. 1989, *Learnability and cognition. The acquisition of argument structure*, MIT Press, Cambridge.

Primus, B. 1998, "The relative order of recipient and patient in the languages of Europe", i Siewierska, A. (red.), *Constituent order in the languages of Europe*, Mouton de Gruyter, Berlin.

Rappaport Hovav, M. og Levin, B. 2008, "The english dative alternation: The case for verb sensitivity", *Journal of Linguistics*, 44, s. 129-167.

Rosenbach, A. 2007, "Animacy and grammatical variation – Findings from English genitive variation", *Lingua*, 118, s. 151-171.

Speas, M.J. 1990, *Phrase Structure in Natural Language*, Kluwer Academic Publishers, Dordrecht.

Tallerman, M. 2005, *Understanding syntax*, 2. utgave, Hodder education, London.

Tekstlaboratoriet 2012, http://www.hf.uio.no/iln/om/organisasjon/tekstlab/ [25.04.2012]

# Vedlegg

Starten på to av de kodede inndatafilene gjengis under. Første linje i hver fil angir hvilke faktornavn R skal bruke. Linjer som begynner med # ignoreres av R. Relevante korpustreff forutgås umiddelbart av en linje på formen 'NP … Kilde' eller 'VP … Kilde' som koder hvordan hver faktor er realisert i treffet, og representerer dermed input til R. Korpustreff som ikke er relevante ignoreres av R siden de mangler en slik forutgående linje.

### Taledata fra Big Brother som kodet i starten av inputfilen til R:

```
real verb class vsense animrec animth defrec defth prorec proth accrec accth lrec lth mod

#Results from search inLoading player...Informants: 20
#bigbrother:


#CWB expression: "([((lemma="gi" %c)])] ;"
#Action :  count download sort collocations annotate show metadata metadata
#  distribution delete hits save hits
#: 282
#Results pages: 1

#              Anette  laga (latter) lost in space åssen sveis ? (latter)
#      (handling)   gi  deg da det er ikke noe godt det skjønner du

#              Anette  ØØ (lydmalende ord) sånn (lydmalende ord) (latter) herlig-
#      nei s- legg *   gi  deg da (uforståelig) slutte (lydmalende ord) ok få
#      ballen (lydmalende ord) nei

#              Monica  nytter ikke å benekte det Ramsy det er bare å   gi  opp og
#      si " ok da " * ja men

#              Roy  (uforståelig) nei e ja (handling) (lydmalende ord) der fikk n
#      han   gitt  * (lydmalende ord) hver jævla gang altså Ø så kommer han

NP gi a gi.a 1 0 1 0 0 0 1 0 1 3 BB
#           Big Brother  fra hagen Ø Big Brother har bestemt seg for å   gi
#      deltagerne en hyggelig overraskelse dersom de fullfører jobben men er

NP gi a gi.a 1 0 1 0 1 0 1 0 1 3 BB
#           Big Brother  Big Brother kaller alle gutta inn i skrifterommet for å
#       gi  dem en liten ekstraoppgave dere får ikke være med kjære

#              Rodney  (latter) Ø " not " nei men er bra ok   gi  gi meg gitaren
#      da Ø så skal jeg skal jeg

NP gi a gi.a 1 0 1 1 1 0 1 0 1 1 BB
#              Rodney  Ø " not " nei men er bra ok gi   gi  meg gitaren da Ø så
#      skal jeg skal jeg bare

NP gi t gi.t 1 0 1 1 1 0 1 1 1 1 BB
#              Rodney  vi dusjer Ø og så går vi inn og så   gir  vi de rosene Ø
#      det trur jeg er best *

#              Anette  du hva jeg skal (uforståelig) nei men vet du hva   gi  vet
#      du hva jeg skal ikke holde lang tale men

#           flere  i tre fire minutter jeg * (uforståelig) var full ennå   gitt
#      (latter) (prating) (uforståelig) jeg så et vinglass og et askebeger

#              Anita  du det er varmt vann her Ø nei ØØ det   ga  seg good
#      morning good morning hej hej hej med så
```

## Skriftdata fra Adresseavisen som kodet i inputfilen til R:

```
real verb class vsense animrec animth defrec defth prorec proth accrec accth lrec lth mod

#Resultater fra søk i tagget bokmålskorpusResultater fra søk i tagget
#bokmålskorpus
#Dato for søk: 06.10.11 12.45

# Søkestreng: [(tagg=".gi\".*" | tagg=".* .gi\".*") & tagg=".* verb.*" &
#tagg!=".* pass.*" & tagg=".* @FV.*" & (src="AV.*" )]
#Søk etter: KWIC-konkordans
#Sortert etter kilde

#43 kilder er valgt.
#Søkt i kilde nr. 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
#25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43
#7624 forekomster funnet.

#Konkordans
#Søkestreng: [(tagg=".gi\".*" | tagg=".* .gi\".*") & tagg=".* verb.*" & tagg!=".*
#pass.*" & tagg=".* @FV.*" & (src="AV.*" )] med 60 tegn på venstre side og 90
#tegn på høyre side.

#      AV Ad96 01: go, slowfox, wienervals og quickstep, sjette i vals. Det ga
#      femteplass sammenlagt, og rankingpoeng for uttak til internasjonale
#      mesterskap. Martin
#      AV Ad96 01: «småfisk». Tre muslimer, tre kroater og en serber, - det gir
#      et sterkt fordreid bilde av skyld og ansvar i etterkrigstidens mest
#      brutale krig i Europa
#      AV Ad96 01: kulose, difteri, kikhoste og stivkrampe. Kampen mot polio gir
#      verdifull erfaring, og styrker det allmenne grunnlaget for arbeidet mot
#      andre smittsomme
#      AV Ad96 01: ter seg til Tenerife | CHRISTER STENSÅS | *én « Bull#s eye »
#      gir en melodi som en enarmet banditt på danskebåten. Mange nok gir en tur
#      til Tenerife. Det
#      AV Ad96 01: en melodi som en enarmet banditt på danskebåten. Mange nok gir
#      en tur til Tenerife. Det er førstepremien i trøndersk mesterskap i
#      compact-dart. Presis
#      AV Ad96 01: føler tilhørighet til sin kirke. Kirken formidler mening og
#      gir livsmot ved høytid og fest. Et brudd fra Statens side innebærer ikke
#      at man bare skrelle
#      AV Ad96 01: pfordres til å ta med seg flagg og bannere på kampen. BOVEY GA
#      SEG | Videomillionæren Grant Bovey, som en stund var favoritt til å overta
#      som eier av N
#      AV Ad96 01: piller og har i innværende sesong spilt for Djurgården. GUD GA
#      GODE RÅD | - Det var som om Gud fader sa til meg at jeg ikke skulle gå ut,
#      sa Ian Goodis
#      AV Ad96 01: tilstrekkelig at Forskningsrådets støtte til en bedrift bare
#      gir effekter i bedriften, i form av økt lønnsomhet, høyere sysselsetting,
#      mer eksport etc.
#      AV Ad96 01: pesielle krav. Prosjektene skal gjennomføres på en måte som
#      gir betydelig samfunnsmessig avkastning. Det kan være ved at prosjektene
#      gjennomføres på uni
#      AV Ad96 01: rammen av det bidrag på 150 mill. kroner som Forskningsrådet
#      gir, vil de 12 største bedriftene betale årlig for ca. 140 mennesker som
#      arbeider med sine d
#for      AV Ad96 01: avlastning), og fordi aktiviteten er av en slik art at den gir
#      ringvirkninger i og dermed avkastning for samfunnet. Noen eksempler :
#      Utvikling av vaksi
NP gi a gi.a 1 0 1 1 1 0 1 0 1 8 Ad
#      AV Ad96 01: en sa Berg at han ville leve på sultegrensen hvis en sponsor
#      ga ham muligheten til å drive med hopping på heltid. - Det er utrolig
#      usunt å pine seg ned
#      AV Ad96 01: har jeg ikke opplevd problemet i gruppen jeg trener. NTB | GIR
#      OPP : Hopperen Øyvind Berg gir seg. Årsaken er all fokuseringen på såkalte
#      slankekontrak
#      AV Ad96 01: i gruppen jeg trener. NTB | GIR OPP : Hopperen Øyvind Berg gir
#      seg. Årsaken er all fokuseringen på såkalte slankekontrakter i norsk
#      hoppsport. DRISTIG
```